QuantMig

Frans Willekens

Simulation of international migration with individual preferences and immigration quota.

Deliverable 2.4

QuantMig has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 870299.

History of changes

Version	Date	Changes
1.0	7 January 2022	Issued for Consortium Review
1.1	28 April 2022	Final version submitted as official deliverable to EC

Suggested citation

Willekens, F (2022) Simulation of international migration with individual preferences and immigration quota. QuantMig Project Deliverable D2.4. Southampton: University of Southampton.

Dissemination level

PU Public

Acknowledgments

We are grateful for valuable comments from Guy Abel, Jakub Bijak, Martin Hinsch and Nico Keilman. This document reflects the authors' view and the Research Executive Agency of the European Commission is not responsible for any use that may be made of the information it contains.

Cover photo: iStockphoto.com/Guenter Guni

Contents

1 Introduction

"Migration is complex and uncertain. To be effective, migration policies need to explicitly acknowledge these two defining features of contemporary mobility" (QuantMig proposal, p. 3). In this paper, uncertainty is dealt with by approaching migration as an outcome of a stochastic process. Complexity is accommodated by viewing the world as a system of regions (countries), in which actors operate. The modelling of an interconnected system of regions requires a multiregional model that incorporates directional migration flows and shed light on the dynamics that results. Population models of systems of interconnected regions have been developed in multiregional demography (see Rogers, 1995 for an introduction). They provide a sound foundation for the comprehension and prediction of migration flows in an interconnected world (Raymer et al., 2018).

In this paper an actor-based multiregional model is proposed. In the literature, an actor-based model is often referred to as an agent-based model. An agent is an entity with the capacity to act on one's own initiative and agency is the manifestation of that capacity. The concept of agency is widely discussed across disciplines and a unique definition does not exists (Bandura, 2006; Hitlin and Johnson, 2015). In this paper, agency is defined as the capacity to act in relative freedom on one's preferences. An agent-based model (ABM) is a microsimulation model with actors that have attributes and agency. Actors usually interact with other actors and their environment. I distinguish two types of actors: individuals and governments. Individuals have the capacity to migrate between countries and regions of the world. That capacity is limited by lack of resources, including social capital, and regulations imposed by governments. Individuals have preferences about where they want to live and work (*location preferences*). They value characteristics of places and assign place utilities accordingly. Place utility is a core concept in behavioural models of migration (Wolpert, 1965). The utility an individual assigns to a place determines the perceived attractiveness of that place. The model brings individuals to life by acknowledging individual preferences, agency, restrictions, rejections, adaptation to conditions beyond one's control, and feelings of satisfaction and dissatisfaction. Agent-based models are used extensively in the study of migration (for reviews, see Klabunde and Willekens, 2016; MacAlpine et al., 2021; Thober et al., 2018; and Hinsch and Bijak, 2022).

Governments monitor the aggregate flow of migrants and may impose immigration quota and use selection of admission criteria to balance the self-selection of migrants. In the paper, the restrictions are limited to immigration quota and a single selection criterion: country of origin. The selection criterion results in immigration quota by country of origin. Individuals who are not satisfied with their current place of residence consider migration. Although the path from desire to action is often long and bumpy (Klabunde et al., 2017), for research purposes the path is usually simplified into two characteristic stages. Carling (2002) proposed the "aspiration/ability model", which distinguished between the wish to migrate and the realization of this wish. Willekens (2021) reviews several simplifications of the path from desire to action. In this paper, a desire to live and work in a particular country is represented by the notion of location preference. The ability to act on one's preference is restricted by immigration quota. If the number of individuals who are dissatisfied with their current residence and prefer to live in a particular country exceeds the immigration quota of that country, not all individuals with a desire to migrate will be able to move to their preferred country. Individuals who do not make it may adapt their preferences resulting in

a stay (by choice) or a move elsewhere. It may also result in a stay by necessity, in which case the dissatisfaction remains. An increase in stayers reduces the level of migration. A move elsewhere is known as substitution effect. Dissatisfied individuals may also look for ways to circumvent the restrictions, e.g. by using unauthorized channels. These effects receive considerable attention in the migration literature (see e.g. Simon, 2019, pp. 27ff; Clemens and Gough, 2018; Barslund et al., 2019). De Haas et al. (2019) show that they limit the effectiveness of migration policies. Note that an individual's ability to migrate to the preferred destination depends not only on immigration quota and admission criteria, but also on the preferences of other individuals in the population. The effects of immigration quota and the hidden interactions between individuals they introduce are made explicit in the paper.

In the paper, a distinction is made between the microstate of a system of regions (or microsystem) and the macrostate of a system (macrosystem). A microstate documents, for each individual in the system, the location and the changes in location. Personal attributes may be added, e.g. individual preferences. A description of the microstate requires that each individual is given a unique identification number (ID). A macrostate describes the system in terms of population characteristics and aggregate migration flows, without reference to individuals. If governments show an interest in individual migrants, their interest concerns the microstate. The distinction between micro- and macrostate is introduced to accommodate actors and actions at two levels of aggregation. Immigration quota impose restrictions on the macrostate of the system of regions; they are generally not aimed at microstates. It means that governments limit the number of immigrants, but are indifferent about who immigrates. If governments are not indifferent, they may use individualized visa as a selection mechanism, in which case they interested in the microstate. The distinction between micro- and macrostate is important for another reason; namely, to determine which aggregate migration flow is most likely, given the individual preferences and the immigration restrictions. Any given macrostate can be produced by different microstates. Possible macrostates are not equally probable, however. Some are more probable than other. The probability of a macrostate depends on the number of microstates that are consistent with a given macrostate. The logarithm of that number is the *entropy* of the macrostate. The most probable macrostate is the macrostate with the highest entropy. If constraints are imposed, such as immigration quota and the requirement that the macrostate reflects individual location preferences, the most probable macrostate is the one that satisfied these constraints and maximizes its entropy. Two slightly different entropy concepts exist. The Boltzmann entropy or configuration entropy relies on combinatorics (the mathematics of combinations and permutations). The Gibbs or Shannon entropy, used in information theory, is rooted in probability theory. The latter is used in this paper[1].

The actor-based multiregional model should satisfy an important requirement: models of microstates and models of macrostates must be consistent. That condition is satisfied when the actor-based model, with its strong microsimulation component, is able to produce the same flows of migrants and the same population distribution as the population-based model. The conditions under which the two types of models produce the same results must be determined. In the context of population projection, Willekens (2011) identifies and discusses the conditions that need to be satisfied to ensure consistency between microsimulation and the cohort-component model. In this paper, a probabilistic formalism is used to integrate micro and macro. The probability model is a

[1] Wilson (1970), who introduced entropy maximization to estimate migration flows between places of origin and destination, used the Boltzmann entropy.

micro-level model; it describes individuals. It also describes groups of individuals who are similar, who differ in a defined way that can be described by a probability distribution, or who collectively meet certain requirements. This fundamental property ensures the integration of micro, e.g. individual agency, and macro, e.g. structural constraints. Immigration restrictions do not apply to individuals but to groups of individuals. They are restrictions individuals must meet collectively. The constraints limit the possible outcomes of the individual-based probability models, and hence the number of microstates. Imposing constraints reduces the number of possible microstates. The effects of the constraints are measured by the change in entropy of the macrostate. In the absence of immigration quota (constraints), the migration flow reflects the individual location preferences. Individuals are free to move to their countries of preference. The individual location preferences (microstate) and the aggregate migration flow (macrostate) have the same information content, which is entropy in information theory[2]. The information contained in the location preferences are transferred without noise to the migration flow. Formulated differently, the actions coincide with the aspirations. Everyone is capable of acting on one's preferences. Immigration quota reduce the freedom of movement. The effect of quota on the individual capabilities to move to one's country of preference is measured by the change in information content. A measure of change in information content, which originated in information theory and is widely used across the disciplines, is the Kullback-Leibler (KL) information divergence. That measure is also used in this paper. A migration flow that satisfies the immigration restrictions and best reflects individual location preferences is the migration flow that carries as much information on the individual location preferences as possible. To obtain that flow, the KL information divergence is minimized. An interesting observation is that the migration flow is also the most probable flow, given the immigration quota and the individual preferences. The migration flow that minimizes the KL information divergence also maximizes the likelihood of the flow given the constraints. The duality between maximum entropy and maximum likelihood is well-established (Good, 1963). It provides a productive approach to bridge the divide between population-level modelling and individual-level modelling. The multinomial distribution is pivotal in bridging the micro and macro perspectives.

The role of individual preferences in shaping aggregate migration flows depend on restrictions imposed on the flows. Restrictions reduce the influence of preferences. The nature of the reduction of in addressed in the paper. In the absence of restrictions, the preferences fully determine the migrant flows, as expected. In the presence of immigration quota, the absolute values of location preferences become irrelevant, but ratios of location preferences, i.e. relative preferences, and ratios of relative preferences become the relevant quantities. A relative preference may be thought of as the preference of one individual relative to that of another individual in the population. Suppose two individuals prefer to move to the same region. One individual has a strong preference and the other a mild preference because she finds other regions attractive too, then, in the presence of immigration quota, the first individual has a higher probability to move to the preferred region than the second individual. It is therefore the relative preference that matters. A general observation is that immigration restrictions affect lower-order dependencies, but leave higher-order dependencies intact. That important observation warrants a discussion of dependence structures in directional migration flow (last subsection of section 2). Individual differences in location preferences have a systematic component, determined by manifest differences between individuals, and a random component due to unobserved differences. The

[2] Recall the two notions of entropy. Boltzmann or configurational entropy is the logarithm of the number of microstates that are consistent with a given macrostate. In information theory, entropy measures the information content of a macrostate. A formal treatment of the two concepts is given in Section 2.1.

random component is often associated with the subjective nature of place utilities individuals attach to places. The model captures the subjectivity of place utilities by adding a random component to place utilities. It turns to the proposed actor-based model into random utility discrete choice model. The link is demonstrated formally in Section 2. Here too, the multinomial distribution plays a pivotal role.

The model needs a procedure that describes how individuals respond to the presence of immigration quota. To accommodate that requirement, a distinction is made between *proposed actions* and *actual actions*. An individual who desires to migrate applies for admission in the preferred country of residence (proposed action). The government decides to accept or reject the application, based on eligibility and admissibility criteria, e.g. the number of applications accepted should not exceed the immigration quota. This simple approach resembles procedures adopted in practice. For instance, each year, millions of people apply for a total of about 55,000 immigrant visa offered in the Diversity Immigrant Visa programme of the United States, also known as the Green Card Lottery. Beneficiaries are randomly selected from applicants who meet the criteria. Other countries, e.g. Canada, use a point system that assigns scores to applicants[3]. By distinguishing proposed and actual actions, complex selection criteria can be accommodated.

A final issue that needs to be clarified is how, in the model, individual location preferences are formed and influence actions. As already mentioned, the path from desire to action can be long and bumpy. Using the theory of planned behaviour (Fishbein and Ajzen , 2010), Klabunde et al. (2017) propose a model of migration that incorporates preference formation and the influence of preferences on actions. In this paper, a different strategy is adopted. The rational is that utility-based preference formation is not an essential component of the model in the paper. I assume that location preferences are revealed by past migration flows. It is essentially the revealed preference theory introduced by Samuelson (1938) in economics to resolve limitations of utility theory discussed at that time. Revealed preferences assume that actual behaviour is indicative of preferences. According to Samuelson, not desires, aspirations or statements about preferences (stated preferences) matter, but the actions that result. Sen (1999) disagrees. In his capability approach, he emphasizes that turning aspirations into actions requires resources many people do not have. If data permit, revealed preferences may be replaced by stated preferences or aspirations, provided they are good predictors of migration, or by elaborate theories of how preferences are formed and motivate actions.

The proposed model is applied to global migration flows. To that end, the world is viewed as a system of countries, grouped into six regions: (1) EU+EFTA+UK, (2) USA and Canada, (3) Latin America and the Caribbean, (4) Africa, (5) Asia, and (6) rest of the world. Individuals migrate between the regions of this multiregional system, but the migration is constrained by immigration quota imposed by governments. The data consists of lifetime migration and recent migration between all countries and territories of the world during the period 1990-2020. The lifetime migration estimates (migrant stocks) are made available by the United Nations. They are based on official statistics on populations country of residence and country of birth, reported in censuses,

[3] See https://www.canada.ca/en/immigration-refugees-citizenship/corporate/publications-manuals/operational-bulletins-manuals/permanent-residence/economic-classes.html for a description of the procedure and Nalbandian (2021) on the algorithms used to facilitate the selection process. In FY2020, 23.2 million individuals applied (https://travel.state.gov/content/dam/visas/Diversity-Visa/DVStatistics/DV-applicant-entrants-by-country-2019-2021.pdf). During the last years, the number of visas issued was considerably less than the number of lottery winners due to administrative factors (https://www.forbes.com/sites/andyjsemotiuk/2021/10/27/winners-of-us-diversity-green-card-lottery-seek-ways-to-immigrate/).

population registers and surveys. From these migrant stock data, Abel (2013, 2018), Azose and Raftery (2019) and other scholars estimated migration flows between all countries of the world during periods of 5 years from 1990 to 2020. Abel and Cohen (2019) and Berlemann et al. (2021) evaluate the estimation methods used by different scholars. Abel and Cohen conclude that the Azose and Raftery method, which combines demographic accounting and the pseudo-Bayesian approach, is the preferred method. The estimates produced by the Abel and Cohen using the Azose-Raftery method are used for this paper. By way of illustration, the model is used to predict the migration flow between the six regions during the period 2015-20 based on (a) immigration quota during that period and (b) the location preferences revealed by the migration flow during the period 1995-2000. Since real data on immigration quota for all countries of the world are missing, hypothetical quota are used. They are based on the number of immigrants during the period 2015-20, for reasons explained in the paper. The use of migration patterns during a past period for predicting current and future migration flows is an established practice in internal migration research and justified by the stability of migration patterns in time. International migration flows exhibit stable patterns too, motivating Bijak (2010, p. 97), Azose et al. (2016), Bijak et al. (2019) and others to view international migration as an autoregressive process based on the "inertia of self-perpetuating migration patterns". Bijak et al. (2019) assessed several time series models of migration and found that the performance of autoregressive models, which assume stationary processes, is acceptable only when migration patterns are stable (p. 477). Abel et al. (2021) showed that, in addition to flows, international migration networks have been remarkably stable over time. The empirical observations are consistent with migration theory (Massey et al., 1993). The theory and the empirical evidence provide the rationale for using past migration flows as indicative of location preferences.

The paper consists of five sections. Section 2 presents the multilevel multiregional probability model. It also presents a method to estimate migrant transition probabilities from revealed location preferences and information on immigration quota. The method is based on information theory. In Section 3, the model disregards individual uniqueness (individuals are not uniquely identified by IDs) and is applied to predict the macrosystem. The model is presented as an extension of the classical multiregional model. The model predicts the population distribution at t+1 from the distribution at t and migrant transition probabilities that are based on location preferences and restricted by immigration quota. The result of the population-level model will be used as a benchmark to validate the actor-based model. In Section 4, individuals are uniquely defined by IDs and the model is used to predict the microsystem. A virtual population is defined and ach member of that population is followed longitudinally during the period from t to t+1. To accommodate the randomness inherent in individual behaviour, the model is reformulated as a random walk in a system of regions. The random walk is biased by individual location preferences and restricted by the presence of immigration quota. A random walk perspective on migration has been proposed previously by a number of authors. The models presented in Sections 3 and 4 are strongly influence by the Schelling (1971) model. The influence of the Schelling model is made explicit in subsection 3.1. Section 5 concludes the paper and lists a number of ways forward. The paper has three annexes. Annex A is a simple illustration of the essentials of entropy maximization. Annex B gives a detailed description of the data used in the paper. Annex C describes how to create a virtual population. The model is programmed in R.

2 Multiregional probability model: probabilistic framework that integrates micro and macro

Multiregional population models describe the distribution of a population in a system of regions. Each member of a population occupies a place of residence. The place can be any geographical unit, from neighbourhood to group of countries. In this paper, place generally refers to a country or a group of countries. The place of residence is also referred to as *location*. Population-based models do not differentiate between similar individuals in a same location. Individual-based model differentiate between individuals; each individual is identified by a name or identification number (ID). The spatial distribution of the population changes in time due to migration, births and deaths. In this paper, births and deaths are disregarded.

The model integrates micro (individual level) and macro (population level) in a single framework. The framework encompasses the notions of microstate and macrostate of a system, referred to in the introduction. More importantly, however, the model is fully consistent with established spatial interaction models of migration. To demonstrate the consistency, a spatial interaction model is presented that incorporates individual variability. It relies on insights obtained by McFadden, who, in the 1970s, showed that the description of individual variability by a particular probability distribution results in a population-level model with desired properties. It was the start of discrete choice modelling under uncertainty.

The section consists of four subsections. The first presents the individual-based model of population distribution at a point in time. In the second, the model is extended to two points in time. The model predicts the number of stayers and migrants, and identifies who stays and who moves. The introduction of place utilities and location preferences turns the model into a discrete choice model. The introduction of immigration quota and other restrictions makes the model more realistic. The third subsection covers the basic principles underlying the estimation of the model. The final subsection zooms in on the origin-destination dependencies exhibited by a migration flow matrix.

A number of concepts are used in this section that are linked to multistate models. Two core concepts are state probability and transition probability. A state probability is the probability that an individual resides in a given location, one of the possible locations. The state probability varies in time. Therefore, it includes a time index. The transition probability is the probability that an individual, who resides in a particular location (i, say) at the beginning of a time interval, resides in a given other place (j, say) at the end of the interval. The transition probability is a conditional probability. Since the transition of interest is change of residence, the transition probability is also referred to as *migrant transition probability* to prevent confusion with the *migration probability*, which is the probability of the event of migration (change of residence) at least once during an interval.

2.1 State occupancies

Consider a population of n unrelated and independent individuals and let k denote a certain individual. Each individual is assigned a unique identifier (ID). Let ${}_kX(t)$ be a random variable denoting the location of individual k at time t, with ${}_kX(t) \in R$ and R the set of possible locations ($R \in \{1, 2, 3, \ldots., r\}$). The number of possible locations is r. Because R is finite, ${}_kX(t)$ is a discrete random variable. ${}_kX(t)$ may be viewed as outcomes of n independent trials at t ($k=1, \ldots, n$) (which leads to the multinomial distribution of ${}_kX(t)$). In many cases, we are not interested in a specific individual, but in randomly selected individual. In that case, the subscript k is omitted. ${}_kX(t)$ has several possible values. The value that is observed at time t is a realization of ${}_kX(t)$ and denoted by ${}_kx(t)$. Several individuals may occupy the same location. To facilitate the counting of individuals, an indicator variable is introduced. Let ${}_kX_i\ (t)$ be an indicator variable that takes on

value 1 if individual k is in location i at time t and 0 otherwise: ${}_kX_i\ (t) = ind({}_kX(t) = i)$. The number of individuals in location i at time t is

$$ {}_+X_i\ (t) = \sum_{k=1}^{n} {}_kX_i\ (t) = \sum_{k=1}^{n} ind\big({}_kX(t) = i\big) \tag{2.1} $$

Since the total population size at time t is fixed at *n(t)*, the probability that the sum $\sum_{i=1}^{r} {}_+X_i\ (t)$ is equal to *n(t)* is one. The (unknown) number of individuals in location i may be expressed as a proportion of the total.

The locations of individuals are unknown, but some aggregate information on the population is available. Let the set $\boldsymbol{n}(t) = \{n_1(t), n_2(t), n_3(t), \dots, n_r(t)\}$ give the distribution of the population in the system at time t. An element $n_i(t)$ is the count of individuals in i at t. The vector $\boldsymbol{n}$ describes the system at the aggregate level, i.e. the *macrostate* of the system. The *microstate* of the system has the location of each individual member of the population. Many different microstates may produce the same macrostate $\boldsymbol{n}$. When two individuals exchange or swap their location, the microstate of the system changes but the macrostate is not affected. The number of individuals that can exchange their location, and hence the number of microstates, depends on the distribution of the population. The number is largest when the population is uniformly distributed in space. If all individuals are concentrated in one location, they cannot exchange locations and the macrostate has a single microstate only. Boltzmann called the number of microstates associated with a given macrostate the *multiplicity* of the macrostate, and denoted it by W. He called the logarithm of W the *entropy* of the macrostate. He used W to determine the most probable macrostate. The most likely macrostate is the one with the highest number of microstates, i.e. which maximizes W and ln W. This definition of entropy is known as configuration entropy. It differs from the entropy concept used in information theory, which is rooted in probability theory (see further). Maximization of the configuration entropy is known as the combinatorial approach to entropy maximization because it relies on combinatorics (the mathematics of combinations and permutations) (see e.g. Niven, 2009). The entropy of a macrostate is a measure of uncertainty because the higher the number of microstates that are consistent with a given macrostate, the higher the uncertainty about which microstate produces the macrostate (for a discussion see Bawden and Robinson, 2015). Entropy maximization is a guiding principle in assigning probabilities to events and transitions. In this paper, entropy maximization is used to determine the most probable migration flow (macrostate).

Let ${}_kp_i(t)$ denote the probability that individual k is in location i at time t: ${}_kp_i(t) = Pr\{{}_kX(t) = x_i = i\} = Pr\{{}_kX_i(t) = 1\}$. In general, different individuals have different probabilities of residing in i at t. The individual probability depends on personal characteristics, situational or contextual factors, and individual life histories. In addition, the probability that k is in i may depend on the location of other individuals, most likely individuals that who are similar to k or related to k. If all individuals have the same probability to be in location i, then ${}_kp_i(t) = p_i(t)$ for all k. The probability that the number of individuals in region i is equal to $n_i(t)$ is $Pr\{{}_+X_i\ (t) = n_i(t)\}$. The expected number of individuals in i at t is $E[{}_+X_i\ ] = p_i(t)\ n(t)$ and the variance is $Var[{}_+X_i\ ] = p_i(t)\ [1 - p_i(t)]\ n(t)$. The index t is omitted for convenience, unless it is needed to prevent ambiguity.

The probability of observing macrostate $\boldsymbol{n}$ is a multinomial distribution with parameters $\{p_1, p_2, \dots, p_r\}$. The multinomial distribution plays a crucial role in the model proposed in this paper. Therefore the distribution is some detail. In particular, the distribution is used to clarify the difference between the Boltzmann entropy and the entropy concept used in information theory. That distinction is important because the estimation of migration flows traditionally relies on the Boltzmann entropy, following Wilson (1970)'s seminal work, whereas more recent approaches are rooted in information theory. Since the two approaches exists side by side and the entropy concept causes much confusion, this section of the paper includes a relative extensive discussion of the entropy concept in and its relation to the multinomial distribution. A historical perspective is considered the best approach to dissolve the confusion. The probability that location 1 has n_1 residents, location 2 n_2 residents, etc., with $\sum_{i=1}^{r} n_i = n$ and n given, is

$$Pr\{ {}_+X_1 = n_1, {}_+X_2 = n_2, \dots, {}_+X_r = n_r\} = \frac{n!}{n_1!\ n_2! \dots n_r!}\, p_1^{n_1} p_2^{n_2} \dots p_r^{n_r} \tag{2.2}$$

The multinomial distribution is $M: n, p_1, p_2, \dots, p_r$ with index n (given) and parameters $p_1, p_2, \dots, p_r$. Since $\sum_{i=1}^{r} p_i = 1$ (follows from fixing the population size), the multinomial distribution has r-1 free parameters. Hence a reference category must be introduced to make sure that parameter values are unique (see further). The multinomial distribution (2.2) consists of two terms. The second term is the probability of a particular microstate that is consistent with macrostate $\boldsymbol{n}$. The first term is the multiplicity of the macrostate, i.e. the number of microstates that are consistent with the macrostate: $W = \frac{n!}{n_1!\, n_2! \dots n_r!}$. The multinomial distribution bridges the micro- and macro-level descriptions of the system of regions.

If all individuals are independent and all locations are equally likely, then $p_i = p = \frac{1}{r}$ for all i and the probability of a particular micro-level configuration is r^{-n}. The number of microstates is r^n and all microstates are equally likely. In another hypothetical situation in which the population size tends to infinity, the observed proportion of the population in location i tends to the true probability: $\frac{n_i}{n} \to p_i$ and $n_i \to np_i$. In that situation, the probability of a microstate that is consistent with macrostate $\boldsymbol{n}$ is $p_{mic(\boldsymbol{n})}$:

$$p_{mic(\boldsymbol{n})} = \prod_{i=1}^{r} {p_i}^{p_i\, n} \tag{2.3}$$

It is the probability of observing a particular spatial distribution of individual IDs or an ordered sequence of individual IDs. The logarithm of a particular micro-level configuration is

$$ln(p_{mic(\boldsymbol{n})}) = n \sum_{i=1}^{r} p_i \log p_i = -n\, H$$

where

$$H = \frac{1}{n} ln \frac{1}{p_{mic(\boldsymbol{n})}} \tag{2.4}$$

and $\frac{1}{p_{mic(\boldsymbol{n})}}$ is the expected number of microstates. It is a function of H: $\frac{1}{p_{mic(\boldsymbol{n})}} = exp[nH]$. H is the entropy concept introduced by Shannon (1948), the founder of information theory. It is the *expected* information content of observing the location of a random individual. The multinomial distribution demonstrates the difference between the Boltzmann entropy and the Shannon

entropy. Boltzmann focused on the first term of the multinomial distribution and assumed that all microstates are equally probable. Shannon entropy emerges from the second term.

Gibbs (1902) extended Bolzmann's work by removing the assumption that all microstates are equally probable. He found that some microstates are more probable than other. He introduced the probability distribution of microstates and defined entropy as a measure of uncertainty in the probability distribution of microstates. The Gibbs approach to entropy became known as the *probabilistic approach*. The combinatorial and the probabilistic approaches exist side by side. The Gibbs entropy concept coincides with that of Shannon. Gibbs concentrated on measuring the level of uncertainty, while Shannon concentrated on the amount of information needed to reduce the uncertainty to an acceptable level. When a system is in equilibrium the Boltzmann entropy agrees with the Gibbs and Shannon entropy (Jaynes, 1965). Jaynes (1957a, 1957b) introduced the maximum entropy principle into statistics and showed that entropy maximization offers a unified framework for statistical inference when data are incomplete and knowledge is limited. He adopted the probabilistic approach to entropy maximization. The approach is consistent with Shannon (1948)'s information theory (Jaynes, 2003). Entropy maximization is a mathematical programming problem of maximizing an objective function subject to one or several constraints. The objective function is not a simple function but a set of functions (sum or integral). The problem is solved using calculus of variation (variational principle), which relies on the method of Lagrange multipliers. For a relatively recent review of the principle of entropy maximization, its spread across disciplines, and its extension to determine most probable paths or trajectories, see Pressé et al. (2013). Entropy maximization is currently the dominant method for estimating migration flows in the presence of different types of prior information (see Section 2.3). The probabilistic approach to entropy is used in this paper. Note that maximization of (2.1) gives maximum likelihood estimates of $p_1, p_2, \ldots, p_r$. Good (1963) proved the duality between maximum entropy and maximum likelihood.

Since the total population size is fixed and $\sum_{i=1}^{r} p_i = 1$, the multinomial distribution has r-1 independent parameters. To make the parameters identifiable, a reference category (reference location) and a normalization constant are needed. Let r' denote the reference category. The odds that a randomly selected individual is in location i rather than in the reference location is $\frac{p_i}{p_{r'}}$. The logarithm of the odds is the logit $\eta_i = \ln\frac{p_i}{p_{r'}} = \ln\frac{p_i}{1-\sum_{j\neq r'} p_j}$. The logit of the reference location is 0, hence $\eta_{r'} = 0$. The probability that a randomly selected individual is in the reference location is $p_{r'} = \frac{1}{\sum_{j=1}^{r} \exp(\eta_j)}$. The probability that the individual is in i is $p_i = \frac{exp[\eta_i]}{\sum_{j=1}^{r} \exp(\eta_j)}$. This are well-known expressions in multinomial logit models and multinomial logistic regression. The denominator is a normalization factor to ensure that the probabilities sum to one. The normalization factor is also known as the partition function, a term first used in statistical mechanics[4]. It represents a statistical ensemble, that is a set of microstates a system can be in. The normalization factor encodes how the total population in the system is partitioned among the different microstates. That explains why the factor is also called partition function.

2.2 Transition probabilities

This section consists of four subsections. First, an individual-based model (IBM) of relocation is presented (subsection a). Change of residence is measured by comparing the place of residence at two points in time. The model is restricted to actual relocations, without any reference to place

[4] In thermodynamics, one over the normalization factor is the Boltzmann factor.

utilities or preferences. This explains the reference to individual-based model and not to actor-based model. In subsection b, preferences and location choices are introduced. Individuals are given some agency. An individual attaches a utility to possible locations and the utility attached determines the individual's preferred location. The place utility an individual assigns to a place depends on (a) observed and unobserved place attributes and (b) the individual's observed and unobserved personal attributes. Place utility is subjective. If several attributes jointly determine the utility individual k attaches to place j, then the expected place utility is used, consistent with the expected utility theory (for details, see Willekens, 2021). The expected utility of a set of places combined is the sum of the utilities individual k attaches to each of the places, weighted by the probability of a relocation to that place. The addition of place utilities and location preferences extends the IBM to a choice model, in which the relocation is the outcome of a discrete choice model. In subsection c, accessibility is introduced: destinations vary in accessibility. Destinations that are near demand less financial resources than destinations that are distant. Cultural differences and differences in language also reduce accessibility. In the fourth subsection, a particular type of accessibility restriction is introduced: the restriction of the number of immigrants. The addition of agency, variation in accessibility and immigration restrictions affect the migrant transition probabilities in particular ways. Note that the proposed model covers all individuals in the population of size n, also the individuals who are satisfied with their current place of residence and have no desire to move.

2.2.1 Migration transition

The probability that individual k is in i at time t and in j at t+1 is the joint probability: $Pr\{{}_kX(t)=i,\ {}_kX(t+1)=j\}$. The joint probability may be expressed as the product of a marginal probability and the associated conditional probability:

$$Pr\{{}_kX(t)=i,\ {}_kX(t+1)=j\} = Pr\{{}_kX(t+1)=j \mid {}_kX(t)=i\}\, Pr\{{}_kX(t)=i\} \tag{2.5}$$

with $Pr\{{}_kX(t)=i\}$ the probability that k is in i at t (marginal probability) and $Pr\{{}_kX(t+1)=j \mid {}_kX(t)=i\}$ the conditional probability that k is in j at t+1, provided k is in i at t. The probability that k is in location i at t is referred to as location probability and denoted by ${}_kp_i(t)$. In probability theory and multistate demography, it is known as the state probability, the probability that individual k occupies state i of the state space $R \in \{1,2,3,\ldots,r\}$. The conditional probability is referred to as *transition probability* and denoted by ${}_kp_{j|i}(t,t+1)$, which is more conveniently written as ${}_kp_{ij}(t,t+1)$. The probability that k is in j at t+1 is

$${}_kp_j(t+1) = \sum_{i=1}^{r} {}_kp_{ij}(t,t+1)\ {}_kp_i(t) \tag{2.6}$$

which is known as state equation. It expresses the state probability at t+1 in terms of the state probabilities at t and the transition probabilities. It is the fundamental equation of multiregional population models (see e.g. Rogers, 1995, Chapter 2). However, it does not apply to the population, but to individual k.

The joint distribution may also be written as the marginal probability that individual k is in j at time t+1 and the conditional probability that k is in i at t, provided k is in j at t+1:

$$Pr\{{}_kX(t)=i,\ {}_kX(t+1)=j\} = Pr\{{}_kX(t)=i \mid {}_kX(t+1)=j\}\, Pr\{{}_kX(t+1)=j\} \tag{2.7}$$

The conditional probability is a *recruitment probability*. It is the probability that individual k who is admitted to j is recruited from i, i.e. originated in i. The recruitment probability is also known as

the admission probability. The recruitment probability may be computed from the transition probability and state probabilities. From (2.5) and (2.7) we have:

$$Pr\{ {}_kX(t) = i \mid {}_kX(t+1) = j\} = Pr\{ {}_kX(t+1) = j \mid {}_kX(t) = i\} \frac{Pr\{ {}_kX(t) = i\}}{Pr\{ {}_kX(t+1) = j\}} \tag{2.8}$$

The expression is similar to the Bayes formula and the derivation is the same, but the interpretation is different. Equation (2.8) can be given a Bayesian interpretation, but that is beyond the scope of this paper. Transition and recruitment probabilities are the main parameters of the multiregional model with immigration constraints.

State probabilities may be combined in the state vector ${}_k\boldsymbol{S}(t)$ with elements ${}_kp_i(t)$ and the transition probabilities in the transition matrix

$${}_k\boldsymbol{P}(t,t+1) = \begin{bmatrix} {}_kp_{11}(t,t+1) & {}_kp_{12}(t,t+1) & {}_kp_{13}(t,t+1) & \cdots & {}_kp_{1r}(t,t+1) \\ {}_kp_{21}(t,t+1) & {}_kp_{22}(t,t+1) & {}_kp_{23}(t,t+1) & \cdots & {}_kp_{2r}(t,t+1) \\ {}_kp_{31}(t,t+1) & {}_kp_{32}(t,t+1) & {}_kp_{33}(t,t+1) & \cdots & {}_kp(t,t+1) \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ {}_kp_{r1}(t,t+1) & {}_kp_{r2}(t,t+1) & {}_kp_{r3}(t,t+1) & \cdots & {}_kp_{rr}(t,t+1) \end{bmatrix} \tag{2.9}$$

The element ${}_kp_{ij}(t,t+1)$ denotes the probability that k who resides in location i at time t resides in j at time t+1. The diagonal element ${}_kp_{ii}(t,t+1)$ is the probability that k, who is in i at t, is also in i at t+1. It does not imply that k stays in i from t to t+1 on a continuous basis. Repeated emigrations and return migrations are allowed. The probability distribution of the possible locations at t+1 is

$$\left[{}_k\boldsymbol{S}(t+1)\right]' = \left[{}_k\boldsymbol{S}(t)\right]' \; {}_k\boldsymbol{P}(t,t+1) \tag{2.10}$$

with $\left[{}_k\boldsymbol{S}(t)\right]'$ is a row vector. This is a matrix expression of the fundamental equation in multiregional modelling, but it is specified at the individual level. Note that the convention in multiregional demography is to write ${}_k\boldsymbol{S}(t)$ as a column vector and to use the transpose of the transition probability matrix (2.9).

The model may be extended by including covariates and contextual factors, i.e. drivers of migration. The transition probability becomes

$${}_kp_{ij}(t,t+1) = Pr\{ {}_kX(t+1) = j \mid {}_kX(t) = i;\; {}_k\Phi(0,t)\}$$

with ${}_k\Phi(0,t)$ the values of the relevant covariates of k and the contextual factors that influence the probability that k, who is in i at t, is in j at t+1.

The expected gross migration flow from i to j is

$$E\left[n_{ij}(t,t+1)\right] = \sum\nolimits_k {}_kp_{ij}(t,t+1) \;\; {}_kp_i(t) \tag{2.11}$$

where summation is over all individuals in the population. In case the individuals are identical: $E\left[n_{ij}(t,t+1)\right] = p_{ij}(t,t+1)\; p_i\;(t)\; n(t)$.

2.2.2 The motives of relocation: place utilities and location preferences

People prefer places for various reasons: employment, to be close to family, safety, etc. The value attached to a place is summarized in the place utility concept (for details, see Willekens 2021). The utility an individual attaches to a places is subjective; it is based on incomplete and often coloured knowledge about places. Furthermore, places have multiple attributes and individuals weight

attributes differently, which gives rise to a utility function. The place utility predicted by the utility function is a multi-attribute utility (Smith, 2010, Chapter 6). In the Schelling model, place is a neighbourhood and the place utility is determined by a single place attribute; namely, the ethnic composition of a neighbourhood. In international migration models, places are countries and the utilities individuals assign to places depend on job opportunities, family or other relationships in the country, physical and social protection, and other place attributes. Individuals differ in what they consider relevant at a given point in time. It depends, for instance, on the stage in the life course. A consequence is that individuals differ in the places they consider plausible destinations (choice set). A consequence of the incomplete knowledge is that the utility an individual attaches to a place is not determined entirely by the objective and observable characteristics of that place. Unobservable or latent place characteristics and subjective factors influence the individual place utility. Place utility is a subjective degree of belief that a place meets desires or aspirations. A place with a high place utility is attractive.

To distinguish the utility associated with observables and the utility associated with unobservables, the place utility consists of two components: a deterministic component and a stochastic component. The stochastic component is in fact a composite measure of multiple uncertainties. One random component may be defined for unobserved differences between individuals, including differences in taste, and another for unobserved differences between places, but that is not done in this paper. Let ${}_kU_j$ denote the utility individual k attaches to place j:

$$ {}_kU_j = {}_kv_j + {}_k\varepsilon_j \quad with\, j \in R \tag{2.12}$$

with R the set of possible locations, ${}_kv_j$the individual-specific deterministic component of utility and ${}_k\varepsilon_j$ the stochastic component. The utility ${}_kU_j$ is a random variable. Although the utility may change in time, time is omitted for convenience. ${}_kv_j$ may be replaced by a utility function in which place utility is a weighted sum of several place attributes and depends on personal attributes. The stochastic component is an idiosyncratic individual-specific term. It is the unobserved utility associated with latent characteristics that influence the attractiveness of j and individual differences in values and taste. For a recent discussion in the context of international migration, see Beine et al. (2021). If the distribution of the random component in a population is known, the value of the random component in k's place utility function, ${}_k\varepsilon_j$, is obtained by a random draw from the distribution. Different distributions are used in random utility theory and discrete choice models. For a recent overview, see Haghani et al. (2021). One distribution is particularly interesting, namely, the Extreme Value distribution or Gumbel distribution (see further).

The absolute value of the utility an individual attaches to a place is not really relevant. What matters is the *relative place utility* (Train, 2009, p. 29; Beine et al., 2021). An individual's preference for a place depends on the subjective utility attached to that place relative to the utility attached to the other places that the individual considers possible places of residence. Relocation is motivated by perceived differences in place utility. Most people attach a relatively high utility to their current place of residence, in part because of the locational capital (Willekens, 2021). They have no motivation to move unless their situation changes drastically.

A common decision rule people use to determine the preferred region of residence is to select the place with the highest place utilities or one of the places with a sufficiently high place utility. If the favourate location is the one with the highest place utility, then the probability that k prefers j over other regions in the choice set is the probability that region j has a higher place utility than any other region in the system of regions:

$$ {}_kp_j = Pr\{ {}_kX = j\} = Pr\{ {}_kU_j > {}_kU_h \; for\; all\; h \neq j\}$$

$$= Pr\{ {}_k v_j + {}_k \varepsilon_j > {}_k v_h + {}_k \varepsilon_h \quad for\ all\ h \neq j\}$$
$$= Pr\{ {}_k \varepsilon_h - {}_k \varepsilon_j < {}_k v_j - {}_k v_h \quad for\ all\ h \neq j\} \tag{2.13}$$

The probability that k selects j is the probability that the difference in utility derived from the unobservables is less than the difference in utility derived from observed characteristics, for all possible places of residence besides i. The unobserved portion of the utility determines the distributions of the utility in a population and therefore the form of the discrete choice model. A (multivariate) normal distribution leads to a probit model (Train, 2009, pp. 111ff). McFadden (1974) showed that the discrete choice model takes the form of a logit model if the unobserved portions of the utility are independent and follow an identical type I extreme value distribution (Gumbel distribution) (see also Train, 2009, pp. 41ff). Independence means that the unobserved portions of the place utility assigned to the different places are unrelated. Extreme value distributions are the limiting distributions for the minimum or the maximum of a large number of independent observations from the same arbitrary distribution. The Gumbel distribution has two parameters: a location parameter α and a scale parameter β>0:

$$Pr\{\varepsilon \leq y\} = exp\left[-exp\left(-\frac{y-\alpha}{\beta}\right)\right] \tag{2.14}$$

The standard Gumbel has location parameter 0 and scale parameter 1.

The difference of two independent extreme value random variates with parameters μ and σ is a logistic variate with location parameter 0 and scale parameter β>0. Let

$$\varepsilon^*_{hj} = \varepsilon_h - \varepsilon_j$$

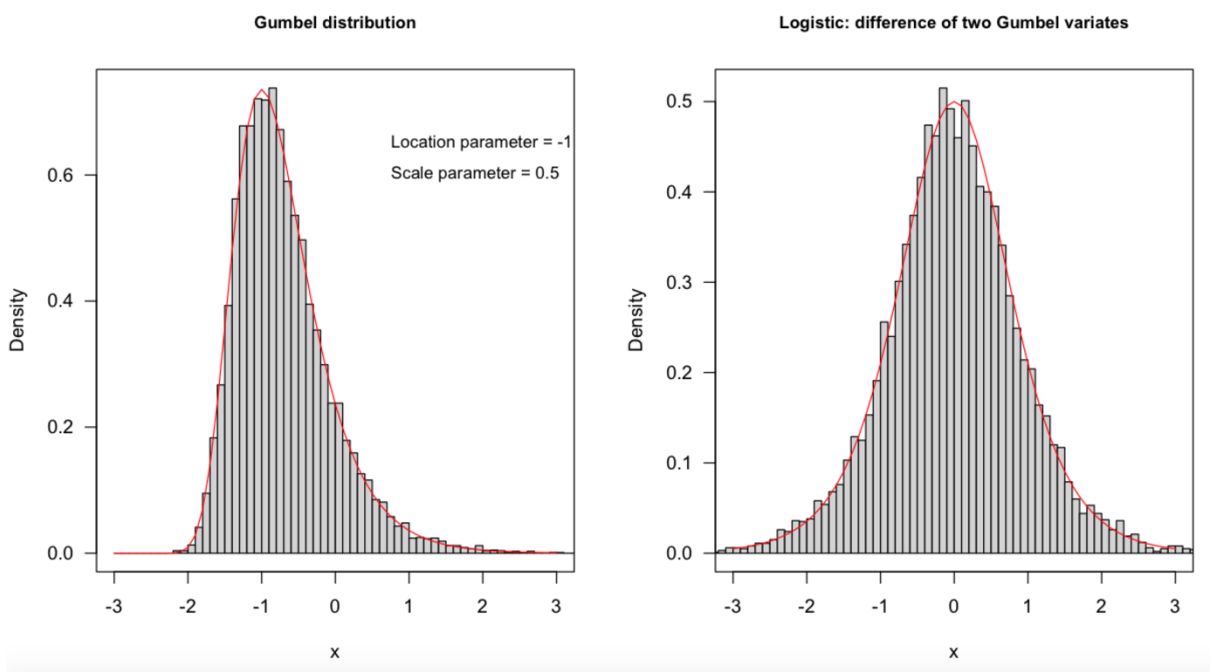
Then ε^*_{hj} follows a logistic distribution:

$$F\left(\varepsilon^*_{hj}\right) = \frac{exp\left[\varepsilon^*_{hj}\right]}{1 + exp\left[\varepsilon^*_{hj}\right]} = \frac{1}{1 + exp\left[-\varepsilon^*_{hj}\right]} \tag{2.15}$$

$F\left(\varepsilon^*_{hj}\right)$ is the probability that the effect of unobservables on the difference in place utilities assigned to regions h and j is less than ε^*_{hj} (for a discussion, see Beine et al., 2021, p. 7). By way of illustration, two random samples were taken from a Gumbel distribution with location parameter α=-1 and scale parameter β=0.5. Each sample is produced by 10,000 random draws from the distribution. The left panel of Figure 2.1 shows the frequency distribution of draws with the overlay of the probability density function of the Gumbel distribution with the same parameters. The right panel of Figure 2.1 shows the distribution of the differences between individual sample values, which is the difference between two Gumbel-distributed random variables. The differences follow a logistic distribution with location parameter 0 and scale parameter equal to the scale parameter of the Gumbel distribution (β=0.5). By way of test, the 10,000 values of the logistic random variable given by $\varepsilon_h - \varepsilon_j$ are used to estimate the parameters of a logistic distribution, using the general-purpose optimization function optim of R and the Nelder-Mead algorithm. The estimated location parameter was -0.00366 and the scale parameter was 0.50726. The estimates are close to the values borrowed from the Gumbel distribution. The difference in the random components of the place utilities individual k attaches to regions h and j may be obtained by drawing a random number from the logistic distribution with location parameter 0 and a scale parameter that is proportional to the assumed variance of the logistic distribution. Note that the variance of the logistic distribution is $\pi^2\ \beta^2\ /\ 3$, with β the scale parameter.

The distribution assumptions of the unobserved quantities determine the form of the discrete choice model and the interpretation of its parameters. The impact of unobserved heterogeneity on the interpretation of the coefficients of logit models and the associated logistic regression models has been addressed by Mood (2010) and Norton and Dowd (2018). The coefficients of the logistic regression models represent not only the effect of a predictor, but confounds the effect of the unobserved heterogeneity.

Figure 2.1 Gumbel distribution and logistic distribution



In the choice model, the probability that individual k prefers region j over any other region is

$$ {}_kp_j = Pr\{\ {}_k\varepsilon_h - {}_k\varepsilon_j < {}_kv_j - {}_kv_h \ \ for\ all\ h \neq j\} $$

$$ {}_kp_j = \frac{exp[\ {}_kv_j]}{\sum_h exp[\ {}_kv_h]} \tag{2.16} $$

The odds that k prefers j is

$$ {}_k\theta_j = \frac{{}_kp_j}{{}_kp_{r'}} = \frac{exp[\ {}_kv_j]}{\sum_h exp[\ {}_kv_h]} \Big/ \frac{exp[\ {}_kv_{r'}]}{\sum_h exp[\ {}_kv_h]} = exp[\ {}_kv_j - {}_kv_{r'}] \tag{2.17} $$

where r' denotes the reference category. The multinomial logit choice model derives its name from the logit expression:

$$ {}_k\eta_j = logit(\ {}_kp_j) = ln\frac{{}_kp_j}{{}_kp_{r'}} = ln\frac{{}_kp_j}{1-\sum_{h\neq r}\ {}_kp_h} = \ {}_kv_j - \ {}_kv_{r'} \tag{2.18} $$

with ${}_k\eta_{r'} = ln\frac{{}_kp_{r'}}{{}_kp_{r'}} = 0.$

The numerator of the odds is the place utility k assigns to region j. The denominator is the sum of the place utilities assigned by k, i.e. k's valuation of the total place utility of the system of regions.

The denominator may be given another behavioural interpretation. Since $\sum_{j=1}^{r} {}_k p_j = 1$, there are $r - 1$ independent parameters of the multinomial distribution of the choice of destination. To make the parameters identifiable, a reference category (reference state) and a normalization constant are needed. The probability that k selects r is ${}_k p_{r'} = \frac{1}{\sum_{h=1}^{r} \exp(v_h)}$. The denominator is a normalization factor. It ensures that the probabilities sum to one:

$$\sum_{j=1}^{r} {}_k p_j = \frac{1}{\sum_{h=1}^{r} exp\left[{}_k v_h \right]} \sum_{j=1}^{r} exp\left[{}_k v_j \right] = 1$$

The probability that k is in region j with place utility ${}_k U_j$ is ${}_k p_j$. The normalization factor encodes how the total utility in the system is partitioned among the different microstates. That explains why the factor is also called partition function. If k is indifferent between the regions, including the current region of residence, then ${}_k v_h = {}_k v$ and $\sum_{h=1}^{r} exp\left[{}_k v_h \right] = r' \, exp\left[{}_k v \right]$ with r the number of regions. Hence the probability that k prefers region j is $1/r'$.

The normalization factor has a simple interpretation in information theory. The logarithm of the normalization factor is the information content of observing an individual (in this case k) who prefers the reference state. The base of the logarithm is the unit of information (Cover and Thomas, 2006): $\ln \sum_{h=1}^{r} exp\left[{}_k v_h \right] = \ln \frac{1}{{}_k p_{r'}} = -\ln {}_k p_{r'}$. The expected information content of knowing individual k's preferred location is the entropy of k's preference distribution[5]:

$$H = \sum_{j=1}^{r} {}_k p_j \; ln \frac{1}{{}_k p_j} = -\sum_{j=1}^{r} {}_k p_j \, ln \, {}_k p_j \tag{2.19}$$

This gives a behavioural interpretation to the entropy concept. The entropy of the place preference distribution is maximum when k is indifferent about the possible destinations, i.e. the place utilities are uniformly distributed. In that case, the entropy is equal to the number of regions. If k is not interested in any region except one, the entropy is zero. A flat distribution of place utilities has a high entropy and a highly skewed distribution a low entropy.

The logit model implies restrictive assumptions and several extensions have been proposed. A characteristic feature of the logit model is that the relative probability of choosing between two alternative destinations depends on the attractiveness of these two options. Other alternatives are irrelevant (McFadden, 1974). It implies that the ratio ${}_k p_j \; / \; {}_k p_{r'}$ does not depend on any alternatives other than j and r'. The odds of choosing j rather than r' is the same no matter what other alternatives are available or what the attributes of the other alternatives are. The independence of the odds from other alternatives is known as the *independence from irrelevant alternatives or IIA property* (Train, 2009, p. 54). While mathematically convenient, this assumption is violated in most contexts where discrete choice models are applied (Train, 2009). Beine et al. (2021) list several extensions that are needed to make the logit model more realistic. First, stayers and intended movers are very different, and foreign destinations are therefore likely to be more correlated with each other than with the domestic destination. Second, some foreign destinations will be more correlated among themselves compared with others because they share characteristics that are unobserved or are not included in the model. The European Union and the Schengen area are cases in point. Other areas aim at a free movement of labour, e.g. the African Union Protocol on Free Movement of Persons, and the Association of Southeast Asian Nations

[5] In information theory, it is a convention that $0 \ln(0) = 0$.

(ASEAN) (Okunade, 2021). Third, random terms are likely to be spatially correlated if migrants rationally decide to limit the acquisition of information to a subset of alternatives, if some of their (ex-post) observed characteristics have been acquired after moving, or if migration influences the distribution of country characteristics. Beine et al (2021) extend the logit choice model to a nested model with overlapping nests. Each respondent faces a large choice set, comprising more than 200 countries worldwide. The nested logit (NL) model requires this choice set to be partitioned into non overlapping nests. The cross-nested logit (CNL) model relies on overlapping nests. They give the same weight to each nest. The probability of choosing a particular destination j can be decomposed into the probability of choosing a particular subset of destinations m and the probability of choosing j within the subset m:

$$P_n(j|C) = \sum_{m=1}^{M} P_n(m|C)P_n(j|m). \tag{2.20}$$

$$P_n(j|C) = \frac{e^{\mu V_{jn}}}{\sum_{j \in C}^{J} e^{\mu V_{jn}}}.$$

The parameters μ_ms capture the similarity between the random terms within nest m. The cross-elasticity for destination j implied by the logit model is the same across all other destinations (i.e., it does not depend on the specificity of location j).

The acquisition of information on possible places of residence is costly and time-consuming. It is therefore rational to economize attention, i.e. to collect and process information that is considered useful and ignore information that is not worth the effort of acquiring and processing[6]. It is known as the *rational inattention* approach to information acquisition. Information about place utilities is incomplete and imperfect (noisy). Individuals cannot collect and process all available information and cannot remove all the noise. However, they can choose which pieces of information to attend to. A rational decision-making process does not require perfect information, but an information acquisition strategy that reduces the uncertainty in the set of place utilities to an acceptable level at least cost. Anas (1983) and Sims (2003) use information theory to describe the level and distribution of uncertainty in the distribution of utility among the alternatives. A measure of uncertainty in the distribution of place utilities is the entropy of the distribution. The acquisition of relevant information on factors determining place utilities leads to a more skewed distribution of place utilities and reduces the number of destinations with a sufficiently high place utility. That facilitates decision making under uncertainty. Matejka and McKay (2015), building on Anas and Sims, distinguish between the utility assigned a priori to an alternative (e.g. place) and the actual utility. The utilities assigned are updated in light of the additional information that is acquired. The authors adopt Bayesian reasoning and add the cost of information acquisition. The aim of information acquisition is to generate subjective place utilities that are close to the actual place utilities. In the authors' view, a high cost of information acquisition motivates an individual to base the ultimate assessment of the utility assigned to the alternatives more heavily on prior beliefs than on (unknown) actual utilities. The amount of uncertainty reduction, measured by the change in entropy of the place utility distribution, depends on the chosen information acquisition strategy. In the *rational inattention* approach to information acquisition, information theory is central to

[6] This type of rationality is known as procedural rationality. It differs from the rationality of the decision itself.

determine an optimal information acquisition strategy (for relatively recent reviews, see Jung et al., 2019, and Maćkowiak et al., 2022). Bertoli et al. (2020) applied this *rational inattention* approach to arrive at a multinomial logit choice model of international migration, using the data presented by Abel (2018).

The odds that individual k prefers region j over the reference region r' usually depends on attributes of k, such as the current place of residence and the stage in the life course. Does knowing more about k contribute to a better prediction of the odds of preferring j over r'? Is it possible to quantify the contribution of additional knowledge? The odds that k prefers j is ${}_k\theta_j = \frac{{}_kp_j}{{}_kp_{r'}} = \frac{Pr\{{}_kD=j\}}{Pr\{{}_kD=r'\}}$, where ${}_kD$ is a random variable denoting the location preference of k. It is the odds in the absence of additional information about k. In other words, individual k is randomly selected from a population. Let's introduce a single attribute of k: current region of residence. The probability that k resides in i and prefers j is the joint probability $\Pr({}_kX = i,\ {}_kD = j)$. If the location preference is independent of the current location, then $\Pr({}_kX = i,\ {}_kD = j) = \Pr({}_kX = i)\ \Pr({}_kD = j)$. The ratio of the joint probability over the product of marginal probabilities is

$$ {}_k\varphi_{ij} = \frac{\Pr({}_kX = i,\ {}_kD = j)}{\Pr({}_kX = i)\ \Pr({}_kD = j)} \tag{2.21} $$

If the ratio differs from one, then ${}_kX$ carries information on ${}_kD$ (and ${}_kD$ carries information on ${}_kX$). The expected value of φ over all possible locations is the mutual information between ${}_kX$ and ${}_kD$ (subscript k omitted for convenience):

$$ E[\ln(\varphi)] = \sum_{i,j} \Pr(X = i, D = j) \ln \frac{\Pr(X = i, D = j)}{\Pr(X = i)\ \Pr(D = j)} \tag{2.22} $$

which may be written as

$$ E[ln(\varphi)] = \sum_{i,j} p_{X,D}(i,j)\ ln \frac{p_{X,D}(i,j)}{p_X(i)p_D(j)} $$

Mutual information is an important concept in information theory (Cover and Thomas, 2006). It is the Kullback-Leibler information divergence with the product of marginal distributions (independence) as the auxiliary distribution (see further). If the mutual information of two random variables is high, i.e. if they share information, than the outcome of one can be predicted from knowledge of the other.

The joint probability may be written as a product of the conditional probability and the marginal probability:

$$ \Pr(X = i, D = j) = Pr\{D = j|X = i\}\ Pr\{X = i\} = \ Pr\{X = i|D = j\}\ Pr\{D = j\} \tag{2.23} $$

Hence

$$ \varphi_{ij} = \frac{\Pr(X = i, D = j)}{Pr\{X = i\}\ \Pr(D = j)} = \frac{Pr\{D = j|X = i\}}{\Pr(D = j)} = \frac{Pr\{X = i|D = j\}}{Pr\{X = i\}} \tag{2.24} $$

$\Pr(D = j)$ is the probability that individual k prefers region j if no information is available on k. $Pr\{D = j|X = i\}$ is the probability that k selects j, provided k is a resident of region i. Note that, if a selection of j is followed by a migration to j, then $Pr\{D = j|X = i\}$ is the probability that a resident of i migrates to j. The conditional probability $Pr\{X = i|D = j\}$ is the probability that an individual who prefers j resides in i. If a selection of j is followed by a migration to j, then $Pr\{X = i|D = j\}$ is an recruitment probability. It is the probability that an individual who prefers region j, randomly

selected from all individuals in a population, "is recruited" from region i or originated from region i. Recruitment probabilities are inverse probabilities because the current region of residence is inferred from data on destination preferences (for a discussion of inverse probabilities in statistical inference, see Fienberg, 2006). Recruitment probabilities play an important role in the model proposed in this paper. They facilitate the accommodation of immigration restrictions. That is the subject of the next section.

If the region of current residence has no effect on an individual's preference, then $\varphi_{ij} = 1$. The larger the deviation from one, the higher the effect of current residence on location preference and the more important is the knowledge of current residence for a prediction of the preferred region of residence. φ_{ij} is the odds that data on current residence improves the prediction of the location preference. The odds may be used to assess the proposition or test the hypothesis that data on current residence are relevant for predicting an individual's preference.

Note that Bayes' theorem follows directly from (2.23) and (2.24):

$$\varphi_{ij}\, Pr\{X = i\} = \frac{Pr\{D = j, X = i\}}{\Pr(D = j)} = \frac{Pr\{D = j|X = i\}\, Pr\{X = i\}}{\Pr(D = j)} = Pr\{X = i|D = j\} \tag{2.25}$$

with $Pr\{X = i|D = j\}$ is the probability that an individual who prefers region j resides in region i.

The odds that individual k prefers j rather than r' is $\theta_j = \frac{p_j}{p_{r'}} = \frac{Pr\{D=j\}}{Pr\{D=r'\}}$. It depends on the relative place utility attached to j. The odds that k prefers j, given the knowledge that k is resident of i is $\frac{Pr\{D=j|X=i\}}{Pr\{D=r'|X=i\}}$. The odds ratio is

$$\frac{\dfrac{Pr\{D = j|X = i\}}{Pr\{D = r'|X = i\}}}{\dfrac{Pr\{D = j\}}{Pr\{D = r'\}}} = \frac{Pr\{X = i|D = j\}}{Pr\{X = i|D = r'\}} \tag{2.26}$$

and

$$\frac{Pr\{D = j|X = i\}}{Pr\{D = r'|X = i\}} = \frac{Pr\{X = i|D = j\}}{Pr\{X = i|D = r'\}}\ \frac{Pr\{D = j\}}{Pr\{D = r'\}} \tag{2.27}$$

The term $\frac{Pr\{D=j\}}{Pr\{D=r'\}}$ may also be interpreted as the odds in favour of the proposition that k prefers j (proposition H1) rather than the alternative proposition that k prefers r' (proposition H2), in the absence of data on the current place of residence. The odds ratio is the relative place utility a resident of i attaches to region j and the relative place utility attached to j by a randomly selected member of the population. The overall population serves as a reference category. In Bayesian statistics, the odds $\frac{Pr\{D=j\}}{Pr\{D=r'\}}$ is called the prior odds, while $\frac{Pr\{D=j|X=i\}}{Pr\{D=r'|X=i\}}$ is the posterior odds (after information on the current place of residence). The factor transforming the prior odds into the posterior odds is the Bayes factor (Kass and Raftery, 1995, p. 776). It is $\frac{Pr\{X=i|D=j\}}{Pr\{X=i|D=r'\}}$, the ratio of probabilities that a particular proposition predicts the data. The Bayes factor, a term coined by Good, is the ratio of the posterior odds to the prior odds. In Bayesian analysis, the Bayes factor is used to assess whether data or evidence supports a proposition, belief, theory or model. The proposition is that individual k prefers region j. The Bayes factor summarizes the evidence provided by the data in favour of the proposition. The methodology for quantifying the evidence in favour of a proposition was developed by Jeffreys (1939). The logarithm of the Bayes factor is commonly referred to as the *weight of evidence* in favour of the proposition that k prefers j. The concept was introduced by Good (1950; 1985, p. 253).

Note that $\frac{Pr\{D=j|X=i\}}{Pr\{D=j\}} = \varphi_{ij}\ Pr\{X=i\}$ and $\frac{Pr\{D=r'|X=i\}}{Pr\{D=r'\}} = \varphi_{ir'}\ Pr\{X=i\}$. Hence the Bayes factor is

$$\frac{Pr\{X=i|D=j\}}{Pr\{X=i|D=r'\}} = \frac{\varphi_{ij}}{\varphi_{ir'}} \tag{2.28}$$

with φ_{ij} given by (2.21). The ratio $\frac{\varphi_{ij}}{\varphi_{ir'}}$ is the probability that an individual who prefers j has the same region of current residence i as individual who prefers reference region r'.

The reference location may be a given location or a hypothetical location, e.g. a location with the average place utility of all locations. Although individuals use different reference locations, it is convenient to use the same reference location for all individuals or for a group of individuals with similar attributes. The relative value of a place utility is the ratio of two place utilities and may be measured by the ratio of two probabilities, the probability of preferring a given location and the probability of preferring the reference location. In other words, the ratio of two place utilities is the odds that an individual prefers a given location rather than the reference location.

2.2.3 Accessibility of destinations and barriers to immigration

The utilities individuals assign to places are relatively stable as long as the characteristics of places that are considered relevant do not change much. It is the consequence of individuals aiming at some consistency in what they want and the reinforcement of individual preferences by other, including the media. Stable preferences lead to consistent choices and stable behavioural patterns. That stability is used in this paper. It is assumed that the place utility an individual assigns to a location is equal to the place utility individuals assigned in previous years. If the preference for a given location changes, it affects the entire distribution of location preferences. For instance, if a location becomes less attractive or less accessible, two effects emerge. First, more people may decide to stay in their current location implying a decline in mobility. Second, some people may choose another destination. The first effect may be called a level effect, while the second is a substitution effect. The effects are analogous to the income and substitution effects distinguished in consumer choice theory to denote the effects triggered by a change in the price of a product. The model presented in this paper quantifies the level and substitution effects of changes in attractiveness of locations. Barriers to immigration have similar effects than changes in attractiveness. The presence of barriers makes a destination de facto less attractive to most people. Barriers have also additional effects, e.g. return migration. Immigrants are less likely to give up their residence when the option to return no longer exists (Czaika and de Haas, 2013). The presence of substitution effects in the case of barriers to immigration was shown by Wissen and Jennissen (2008), Simon (2018, 2019) and others. Simon (2019, pp. 27ff) distinguishes two substitution effects: a location (destination) effect and a channel effect. The latter is a shift from legal migration to clandestine or unauthorized migration (see also Clemens and Gough, 2018; Barslund et al., 2019). De Haas et al. (2019) show that they limit the effectiveness of migration policies.

In this subsection, two extensions are introduced. First, the presence of obstacles imply that places differ in accessibility. An origin-destination specific measure of accessibility is added. The introduction of an accessibility term has a very interesting side effect: the random utility discrete choice model closely resembles the gravity model. Second, barriers to immigration are introduced.

a. Accessibility

The gravity model and its extensions include a term to measure the effect of accessibility, intervening opportunities or spatial friction, e.g. distance. Ortega and Peri (2013) and Beine et al. (2016) add an accessibility term to discrete choice model in order to capture the effects of obstacles and individual differences in capabilities to remove the obstacles and make it to the preferred location. The utility individual k attaches to region j is

$$ {}_kU_j = {}_kv_j - {}_kc_j + {}_k\varepsilon_j \tag{2.29} $$

with ${}_kc_j$ the disutility or reduction in utility associated with a less than perfect accessibility of j. The term captures effects of obstacles, including distance, costs, differences in culture and language between places, and other forms of spatial friction. The authors show that the introduction of an accessibility term produces a random utility discrete choice model that closely resembles the gravity model. Assume that all potential migrants in the system assign the same place utility to region j: ${}_kv_j = v_j$ for all k. Assume further that accessibility of j varies by region of origin, and does not vary between individuals in that region of origin. In other words, individual differences in preferences for region j are fully determined by differences in access to j. Assume finally that the random term follows an independent and identically distributed type I extreme value distribution. The choice model becomes (omit k)

$$ p_{ij} = \frac{exp[v_j - c_{ij}]}{\sum_h exp[v_h - c_{ih}]} = \frac{exp[v_j]}{\sum_h exp[-c_{ih}]\, exp[v_h]} exp[-c_{ij}] = a_i\, b_j\, exp[-c_{ij}] \tag{2.30} $$

with $a_i = 1/\sum_h exp[-c_{ih}]\, exp[v_h]$, $b_j = exp[v_j]$ and $exp[-c_{ij}] \leq 1$. The last term of equation (2.30) resembles the gravity model. The exponential of the utility reduction due to an imperfect accessibility of j may be written as $c^*_{ij} = exp[-c_{ij}]$. It represents the effects of obstacles to migration from i to j. The model that results resembles the biproportional adjustment (RAS) model, widely used in migration analysis (see e.g. Willekens, 2016):

$$ p_{ij} = a_i\, b_j\, c^*_{ij} \tag{2.31} $$

Notice that

$$ \frac{\partial a_i}{\partial c^*_{ih}} = exp[v_h] = b_j \tag{2.32} $$

which implies that a reduction in the accessibility of destination h for individuals in i, leads to an increase in the preference for j. It is the substitution effect associated with a less than perfect accessibility of regions. If two countries i and j form a union with freedom of movement or reduced mobility constraints, then the mobility between these countries will increase. The odds that a resident of i migrates to j rather than h is

$$ \frac{p_{ij}}{p_{ih}} = \frac{b_j\, c^*_{ij}}{b_h\, c^*_{ih}} $$

It depends on a resident of i's perception of the relative attractiveness of the two regions and their relative accessibility. The odds of staying in i with a change in accessibility of j is

$$\frac{p_{ii}}{p_{ij}} = \frac{b_i}{b_j\, c_{ij}^*} \tag{2.33}$$

If the accessibility of j declines, the odds of staying in i increases, irrespective of changes in attractiveness of alternative destination. It is a consequence of the IIA property of the logit choice model, which is associated with the distributional assumption of the random component of the utility function.

b. Barriers to immigration

Barriers to immigration introduce competition between potential migrants. In the presence of barriers, individual aspirations (preferences) and capabilities to remove obstacles are not enough to make it to a preferred destination. The preferences and capabilities of others are important too. Relative place utilities and relative destination preferences, expressed as odds, are guiding principles in the absence of immigration barriers. In the presence of barriers, *odds ratios* become important. Odds ratios relate relative preferences to background factors, such as current region of residence and personal factors. Background factors affect the strength of relative preferences, and consequently the substitution patterns. A consequence of quota is that relative location preferences (odds) become irrelevant for predictive purposes, but the ratio of relative preferences of any two individuals become important.

Empirical studies of time series of migration flows indicate that individual differences in strength of relative preferences are quite stable in time. That finding has been known for quite some time in internal migration (see e.g. Snickers and Weibull, 1977; Willekens, 1982). Abel and Sander (2014) discovered similar stable dependence structures in international migration. Barthel and Neumayer (2015) identified stable dependence structures in asylum flows. That finding has major consequences for the estimation and prediction of migration flows because the time-invariance or stability is useful knowledge that can improve estimations and prediction. The useful knowledge is usually incorporated in the models of migration in the form of auxiliary migration flow data, usually a historical migration flow that carries information on individual differences in strength of relative preferences. Interestingly, the approach is consistent with iterative proportional fitting (IPF) and the biproportional adjustment method (RAS), which are most popular methods to predict migration flows from incomplete data. These methods have the interesting property that they preserve differences in strength of relative preferences, measured by the odds ratio. The property is known as the i*nvariance of association* principle. The principle is used to impose dependence structures onto migration flows when the available data provide no information on the structure (Rogers et al., 2003; Raymer, 2007; Raymer et al., 2019).

2.3 Estimation of transition probabilities

Logistic regression is the dominant method for estimating transition probabilities when data are complete. The method is inadequate when some necessary data are missing. A number of approaches have been proposed to estimate transition probabilities from incomplete data. One approach is to specify a model of the complete data and to estimate its parameters from the data that are available. The model describes the missing-data-generating mechanism. Little and Rubin (2020[1987]) introduced the Expectation-Maximization (EM) algorithm, which maximizes the likelihood of the data given the model, i.e. it is a maximum-likelihood method. Little and Rubin acknowledge that their approach depends on the specification of the model and they warn for model misspecification. The technique has been applied for estimating migration flows (Willekens,

1999; Abel, 2013). A method that does not require specification of the likelihood function is Approximate Bayesian Computation (ABC) (Beaumont, 2019). The likelihood-free estimation method defines a set of possible parameter values (parameter space) and simulates realizations of the model's outcome under different parameter values. The desired parameter values produce outcomes that are "close" to observed outcomes. A third approach relies on information theory. That approach is used in this paper. No model of the missing data is specified and, by implication, the method is likelihood-free. Missing data are estimated from available data using constrained optimization techniques and the objective function in the optimization problem is derived from information theory. The available data on migration act as constraints and the estimated or predicted migration flows must satisfy these constraints. In other words, the missing data must be consistent with what is known about the data. The approach may be viewed as an imputation of missing data from available data. Imputations are not unique since many alternative imputations satisfy the constraints imposed by the available knowledge. Therefore, the most probable values of a migration flow that are consistent with the known data (imposed as constraints) need to be determined. The most probable values leave a maximum of uncertainty allowed by the available knowledge. They are obtained by maximizing the entropy of the migration flow (macrostate). The principle of entropy maximization (MAXENT) is the modern version of the principle of insufficient reason advanced by Jacob Bernoulli (1654-1705) and the indifference principle advanced by Keynes' (1921, pp. 52-53) (see also Robert, 2011). It states that if we do not have sufficient reason to regard one outcome more probable than another, we should treat them as equally probable. The maximum entropy distribution leaves a maximum uncertainty and is therefore least informative, while accounting for the available knowledge (Jaynes, 2003)[7]. Jaynes (1957a, 1957b), who introduced entropy maximization in statistics, showed that entropy maximization offers a unified framework for statistical inference when data are incomplete and knowledge is limited. Wilson (1970) introduced entropy maximization in the study of migration, but used the combinatorial approach to entropy maximization. Jaynes adopted the probabilistic approach. The latter approach is followed in this paper.

The estimation of true migration flows may benefit from auxiliary information. Auxiliary data contribute information that is not supplied by data on the true flows. Estimation methods give priority to information on the true flows and use the auxiliary information to fill in the missing pieces. The most probable estimates of the true flows are those that extract as much information from the available knowledge as possible and acknowledges the uncertainty that remains. An indicator of success is the information content of the estimates. It should be as close as possible to the information content of the available data. In information theory, the difference in information content is measured by the Kullback-Leibler (1951) (KL) information divergence (also known as cross-entropy, Rubeinstein and Kroese, 2016). The KL information divergence and the ABC algorithm have much in common. Spiliopoulos (2020) discusses the ABC algorithm through the lens of information theory.

Consider a system of regions $R = \{x_1, x_2, \dots, x_r\}$, which is denoted by the indices $R = \{1, 2, \dots, r\}$ for convenience. Let ${}_kX(t)$ and ${}_kX(t+1)$ denote the region of residence of individual k at time t and t+1, respectively. The probability that k resides in i at t is the state probability ${}_kp_i(t)$ (see Section 2.2). The probability that k resides in region i at t and region j at t+1 is the joint probability $Pr\{{}_kX(t) = i,\ {}_kX(t+1) = j\} = {}_kp_{ij}\ {}_kp_i(t)$. In what follows, k refers to an individual selected at random from the entire population or from the population of a region, and the subscript k is

[7] Note that Bayes introduced a uniform prior distribution and Jeffreys designed method for producing uninformative priors to reflect ignorance (Fienberg, 2006). That approach also leaves a maximum of uncertainty.

omitted. Recall that $n_i(t)$ denotes the count of individuals in region i and $n_{ij}(t, t+1)$ the number of individuals in region i at t and in region j at t+1. It is common to display count data in a contingency table. Viewing a migration flow matrix as a contingency table of counts has the advantage that methods for the analysis of contingency tables can be used for the estimation of migration flows (Willekens, 1980, 1982). For recent applications to international migration, see Abel (2013, 2017) and Raymer et al. (2019). In what follows, counts may be replaced by proportions using the equality $n_{ij} = \hat{p}_{ij}\, n$ with $\hat{p}_{ij}$ the proportion or relative frequency. As n tends to infinity, the proportion tends to a probability (asymptotically).

Data on the true distribution of migration flows are incorporated as constraints, while the auxiliary data are used as a prior distribution in Bayesian parlance. The two types of information are discussed in some detail. For illustrative purpose, a most simple case is considered first. Suppose that nothing is known about the true joint distribution of places of residence at t and t+1 except the size of the population. In the absence of auxiliary information, the probability that a randomly selected member of the population is in region i at t and in region j at t+1 is $\hat{p}_{ij} = \frac{1}{r^2}$ when the population size n is sufficiently large. The estimated count of persons in i at t and j at t+1 is $\hat{n}_{ij} = n\,\hat{p}_{ij} = \frac{n}{r^2}$. This simple estimation method maximizes the entropy of the cross-classification of individuals by region of residence at t and residence of residence at t+1 (Annex A). Although auxiliary data are absent, they may be imagined as being present but not informative. The least informative distribution is the uniform distribution. Let p_{ij}^0 denote the auxiliary or reference distribution and assume that it is the uniform distribution: $p_{ij}^0 = \frac{1}{r^2}$. The estimates obtained in the absence of an auxiliary distribution are identical to those obtained with a uniform reference distribution.

Let's consider additional information. Suppose the population distribution at t and the distribution at t+1 are known, but the joint distribution is not known. In other words, the marginal distributions of ${}_+X(t)$ and ${}_+X(t+1)$ are known, but their dependence structure is unknown. Information on the true joint distribution is incorporated in the estimation method as constraints: the joint distribution must satisfy given marginal distributions. The missing information, i.e. the dependence structure, is extracted ('borrowed') from the auxiliary distribution. As a consequence, the true distribution and the auxiliary distribution exhibit the same interaction effects. The principle of borrowing effects from auxiliary distributions is fundamental for estimating joint distributions from incomplete data. Good (1963) uses that principle to estimate entries of a contingency table (joint distribution) from marginal totals (marginal distributions) and an auxiliary distribution. He implements the approach by minimizing the Kullback-Leibler divergence (relative entropy) between the true and auxiliary distributions (Good, 1963, p. 912) subject to constraints representing the knowledge about the true distribution. He calls the principle the *principle of minimal discriminability* and sees it as a generalization of the Jaynes' principle of maximum entropy. Good also proves the important duality between minimization of the Kullback-Leibler divergence and Fischer's maximum likelihood method. His duality theorem states that the maximum likelihood estimates of the parameters of the distribution of a discrete random variable (probabilities) are equal to the maximum entropy estimates (Good,1963, p. 927). Earlier he had shown the connection between information theory and Fisher's notion of sufficient statistic (Good, 1956, p. 201). As a consequence of the duality theorem, entropy maximization or minimization of relative entropy (KL divergence) may be used to obtain maximum likelihood estimates. Lusem and Teboulle (1992) give a mathematical description of the duality. Good's principle of minimum

information is the same as the minimum information divergence proposed by Ireland and Kullback (1968) and Kullback (1968)[8].

The entropy maximization problem is to predict the true distribution p_{ij} from (a) the marginal distributions p_{i+} and p_{+j}, and (b) the auxiliary distribution q_{ij}. An alternative, but equivalent formulation is to predict the true migration flows n_{ij} from (a) the known number of individuals in region i at t who are still present in the multiregional system at t+1, denoted by O_i, and the known number in j at t+1 who were also present in the system at t, denoted by D_j. and (b) the destination preferences, which act as auxiliary data. Let n_{ij}^0 denote the number of individuals in region i who prefer to be in region j. The probability that an individual in i prefers j, denoted by p_{ij}^0, is estimated by $n_{ij}^0 / \sum_j \ n_{ij}^0$. Notice that individuals in i may desire to stay in i. In case individuals who desire to stay are excluded: $\hat{p}_{ij}^0 = n_{ij}^0 / \sum_{j \neq i} n_{ij}^0$. Note that, because only individuals present at t and t+1 are included, $\sum_i \ O_i = \sum_j \ D_j$. The number of individuals who prefer region j may exceed the capacity of j or the immigration quota imposed by j. In that case, several individuals cannot move to their preferred destination. They may decide to stay in their current region of residence, move to another suitable destination, or opt for a different channel of migration. Let n_{ij}^* denote the predicted migration from i to j.

The optimization problem is to find the values of n_{ij}^* that reflect the preferences as good as possible, subject to information available on the true flow n_{ij}:

$$\text{minimize } D_{KL}(n_{ij}^* \| n_{ij}^0) = \sum\nolimits_{ij} n_{ij}^* \ ln \frac{n_{ij}^*}{n_{ij}^0} \tag{2.34}$$

subject to

$$\sum\nolimits_j \ n_{ij}^* = \sum\nolimits_j \ n_{ij} = O_i$$

$$\sum\nolimits_i \ n_{ij}^* = \sum\nolimits_i \ n_{ij} = D_j$$

The objective function is the discrete analogue of a functional, an integral of functions. Discrete calculus of variations is used to find the solution to the constrained optimization problem.

To obtain the most probable values of n_{ij}^*, the Lagrangian function is constructed and minimized. The Lagrangian is

$$L = \sum\nolimits_{i,j} n_{ij}^* \ln \frac{n_{ij}^*}{n_{ij}^0} + \sum\nolimits_i \ \lambda_i \left(\sum\nolimits_j \ n_{ij}^* - O_i \right) + \sum\nolimits_j \ \mu_j \left(\sum\nolimits_i \ n_{ij}^* - D_j \right) \tag{2.35}$$

with λ_i and μ_j Lagrange multipliers associated with the constraints. The multiplier λ_i measures the impact of a small change in O_i on L: $\frac{\partial L}{\partial O_i} = \lambda_i$. It quantifies the impact on $D_{KL}(n_{ij}^* \| n_{ij}^0)$ of a change in the number of residents of i who either stay in i or migrate to another region. A decrease in the KL information divergence means that the true migration pattern better reflects the preferences of

[8] Akaike (1973) found that the KL information divergence is the expected value of the log likelihood ratio in favour of the true distribution against the other candidate distributions (for a discussion of the relation between KL divergence and the log-likelihood ratio, see Etz, 2019 and Eshima, 2020). The KL divergence is the core of the Akaike information criterion (AIC). It is also a core component of the generalizations of the Deming and Stephan's IPF method by Ireland and Kullback (1968) and Darroch and Ratcliffe (1972). Darroch and Ratcliffe start from Deming-Stephan 1940, who minimize chi-square.

individuals in the system of regions. The Lagrange multiplier $\frac{\partial L}{\partial D_j} = \mu_j$ measures the effect on the KL information divergence of a change in the number of people admitted to region j. If a decrease in D_j causes an increase in $D_{KL}(n_{ij}^* \| n_{ij}^0)$, then a reduction in the admission of immigrants (or immigration quota) produces a migration flow that reflects the preferences less than before the reduction of the quota. The interpretation of Lagrange multipliers has received very limited attention in migration modelling. Nagurney et al. (2020, 2021) use network equilibrium models and the theory of variational inequalities (discrete calculus of variations) and Lagrangeans to gain insights as to the impacts of regulations, including immigration quota, on place utilities for different classes of migrants, and on international refugee migration flows. Notice that immigration quota and other policy measures enforced by governments are captured via the constraints on migration flows. The constraints therefore may reflect different types of knowledge about the true migration flows.

The necessary conditions for the minimum are

$$\frac{\partial L}{\partial n_{ij}^*} = 0 = \ln \frac{n_{ij}^*}{n_{ij}^0} + 1 + \lambda_i + \mu_j \tag{2.36}$$

$$\frac{\partial L}{\partial \lambda_i} = 0 = \sum\nolimits_{j=1}^{r} n_{ij}^* - O_i$$

$$\frac{\partial L}{\partial \mu_j} = 0 = \sum\nolimits_{i=1}^{r} n_{ij}^* - D_j$$

The first equation yields the desired distribution in terms of the Lagrange multipliers:

$$\frac{n_{ij}^*}{n_{ij}^0} = exp\left[-\left(1 + \lambda_i + \mu_j\right)\right] \tag{2.37}$$

$$n_{ij}^* = exp\left[-(1 + \lambda_i + \mu_j)\right] n_{ij}^0$$

The most likely pattern of relocation cannot be determined analytically. It must be obtained by iteration. The iterative algorithm is a variant of iterative proportional fitting (IPF). Let $a_i = \exp(-1 - \lambda_i)$ and $b_j = \exp\left(-\mu_j\right)$. We may write

$$n_{ij}^* = a_i \, b_j \, n_{ij}^0 \tag{2.38}$$

with a_i and b_j to be determined from the marginal totals using an iterative procedure:

$$a_i \sum\nolimits_{j=1}^{r} n_{ij}^0 \, b_j = O_i$$

$$a_i = \frac{O_i}{\sum_{j=1}^{r} b_j \, n_{ij}^0} \tag{2.39}$$

and similarly

$$b_j = \frac{D_j}{\sum_{i=1}^{r} a_i \, n_{ij}^0} \tag{2.40}$$

The iteration starts with any value of a_i or b_j.

Entropy maximization, and more particularly the combinatorial approach, is the dominant method for the estimation of migration flows by origin and destination. Inspired by Wilson (1970), Chilton and Poet (1973), Willekens (1977, 1982, 1999) and many others adopted the approach. The method

gives the same result as the probabilistic approach based on information theory and presented above. The information theoretic method has also been popular (Snickars and Weibull, 1977; Plane, 1982; Roy and Flood, 1992).

In statistics, the method of estimating the entries of a contingency table from given marginal totals and an auxiliary contingency table is known as iterative proportional fitting (IPF). In economics and geography the method is known as biproportional adjustment and RAS method. IPF is associated with Deming and Stephan (1940) although the authors did not maximize the entropy function but a chi-square-type distance norm. Deming and Stephan's justification for using least-square optimization turned out to be wrong, as noted by Stephan (1942) (see also Fienberg, 1970). Jaynes (1957a, p. 623) notices that the least squares method has some properties of Shannon's information method, but with limitations. Jaynes (2003, p. 346) gave a simple example to illustrate the problem. Entropy maximization produces nonnegative estimates because of the log-transformation. Idel (2016) gives an extensive review of over 70 years of iterative proportional fitting from a mathematical perspective but excludes the Deming-Stephan algorithm. The author also shows that the biproportional adjustment method is the dual problem of the Kullback-Leibler information minimization (Idel, 2016, p. 16). Fienberg (1970), Ireland and Kullback (1968) and Zaloznik (2011) offer good discussions of the algorithm.

2.4 A note on dependency structures and log-linear models

2.4.1 Dependence structures

The constrained minimization of the information divergence between the unknown true migration flows and the individual preferences produces predictions with interesting properties. First, the predicted values n_{ij}^{*} agree with the available information on the true flows (in this case O_i and D_j for all i and j). The information on the true flows is imposed onto the estimates through the constraints. In other words, the bivariate distribution agrees with the marginal distributions. Second, the information not supplied by the constraints is obtained from the auxiliary data, in this case individual preferences, aggregated in counts of individuals by origin and preferred destination (n_{ij}^{0}). It means that the predicted migration flows reflect the individual preferences, but only partly due to the constraints imposed onto the flows.

It is useful to distinguish local and global measures of dependence. The odds ratio or cross-product ratio is a measure of local dependence. It is the ratio of the relative location preferences of a resident of one region and the relative location preference of a resident of another region. Another local measure of dependence is the ratio of the migration between two regions and the migration expected under conditions of independence. It is $\frac{n_{ij}}{n_{i+}n_{+j}/n_{++}}$. Mutual information is a measure of global dependence. It is given in equation (2.22). The global indicator summarizes the effect of local dependencies. If the true migration flow matrix and the auxiliary migration flow matrix, e.g. historical migration flow, exhibit the same dependence structure, then the mutual information is the same in the two matrices and the odds ratios are the same. If no historical migration flow is available, odds ratios may be elicited using other methods, such as expert opinions, and used to impose an association structure onto the estimates (Coffey et al., 2020).

Consider odds ratios in more detail and suppose that the auxiliary data consists of a matrix of location preferences which vary by current region of residence. The odds ratio is given in equation (2.26). If the data consist of population counts, then the odds ratio may be written as:

$$\alpha_{ij} = \frac{\frac{n_{ij}}{n_{iJ}}}{\frac{n_{Ij}}{n_{IJ}}} = \frac{n_{ij}\, n_{IJ}}{n_{iJ}\, n_{Ij}} \tag{2.41}$$

where I and J denote the regions of current residence and preferred residence that act as reference categories in the definition of the odds. Two reference categories are used to accommodate two constraints, the first that the number of people by current region of residence is *n*, the second that the number of people by preferred region of residence is *n* too. A same region may be used as reference category in current residence and preferred residence. In (2.26), we used a single reference category, i.e. $I = J = r'$. The odds ratio is the odds that a resident of region i prefers region j rather than the reference region J, divided by the odds that a resident of the reference region I prefers j rather than J. In other words, it is the relative location preference of a resident of i divided by the relative location preference of a resident of reference region I. If the current region of residence has no influence on the location preference, the odds ratio is one. A large odds ratio centered around the cell (i,j) implies that residents of i have a much higher preference for j than one may expect on the basis of the outflows from i and inflows in j. It measures the extent at which a proposition is supported by the data. The logarithm of the odds ratio α_{ij} can be written as the difference between two linear contrasts:

$$ln\, \alpha_{ij} = ln\frac{n_{ij}}{n_{iJ}} - ln\frac{n_{Ij}}{n_{IJ}} = ln\frac{p_{ij}}{p_{iJ}} - ln\frac{p_{Ij}}{p_{IJ}} = logit(p_{ij}) - logit(p_{Ij})$$

with $p_{ij} = \frac{n_{ij}}{n_{++}}$. Note that the $logit(p_{ij})$ is defined for current region of residence i and $logit(p_{Ij})$ is defined for the reference region I of current residence.

Note that the Bayes factor is an odds ratio too (Kass and Raftery, 1995). The logarithm of the Bayes factor is the weight of evidence (see Section 2.2.b).

Some authors distinguish between nominal odds ratios and local odds ratios (see e.g. Kateri, 2014, pp. 40ff). Nominal odds ratios are defined with respect to a fixed reference cell (I,J), as in equation (2.41). Local odds ratios are computed for a 2 by 2 subtable formed by the entries *i*, *i+1*, *j* and *j+1*:

$$\alpha_{ij}^{loc} = \frac{\frac{n_{ij}}{n_{i,j+1}}}{\frac{n_{i+1,j}}{n_{i+1,j+1}}} = \frac{n_{ij}\, n_{i+1,j+1}}{n_{i,j+1}\, n_{i+1,j}} \tag{2.42}$$

The odds ratio is an appealing measure of association in cross-classified data because it has useful properties. The odds ratio is invariant under the interchange of rows and columns and is invariant under row and column multiplication (Mosteller, 1968, p. 4). The invariance means that the odds ratio does not change if we multiply counts by row factors and column factors. Hence a contingency table of elements

$$n_{ij}^{*} = a_i\, b_j\, n_{ij}^{0} \tag{2.43}$$

exhibit the same odds ratios as contingency table n_{ij}^{0}, for any value of a_i and b_j. That property is known as *multiplicative invariance of association* (Mosteller, 1968). No other measure of association has that property (Bishop et al., 1975, p. 392). The invariance property implies that the odds ratio is independent of marginal totals. It is a margin-free measure of association (Bishop et al., 1975, p. 375).

The IPF preserves the association between cross-classified random variables, something which Deming and Stephan (1940) did not realize (Mosteller, 1968, p. 10). The property of multiplicative invariance of association is particularly interesting for the estimation and prediction of migration flows (Willekens, 1983, p. 192) and has been used frequently in applied research. Willekens (1994; 2016, p. 235) reviews migration estimation methods that incorporate the property, although often not explicit or under a different name. Rogers et al. (2003, 2010) used it to impose spatial structures onto estimates of migration flows. Today, the IPF is the main method for estimating international migration flows by country of origin and country of destination (Abel, 2017; Azose and Raftery, 2019; Raymer et al., 2019).

The preservation of local dependencies implies preservation of global dependence. Equation (2.43) may be written in matrix terms:

$$\boldsymbol{n}^* = \boldsymbol{a}\,\boldsymbol{n}^0\,\boldsymbol{b} \tag{2.44}$$

with $\boldsymbol{a}$ and $\boldsymbol{b}$ diagonal matrices of elements a_i and b_j, respectively.

The matrices $\boldsymbol{n}^*$and $\boldsymbol{n}^0$ in equation (2.44) have equal mutual information, which differs from the mutual information in the true flow matrix $\boldsymbol{n}$**.** The two matrices have equal mutual information because they share local dependencies (odds ratios). The invariance property is a logical step to more general dependence structures in multivariate distributions and copula theory. Copula theory deals with dependence structures in multivariate distributions. In a 3-page note, Sklar (1959) showed that any multivariate joint distribution can be written in terms of univariate marginal functions and a distribution that describes the dependence structure between random variables, known as *copula*. Sklar's theorem is the foundation of copula theory and copula models (see e.g. Nelson, 2006; Czado, Geenens, 2020, Genest, 2021). Copulas are a powerful and flexible tool for modelling associations in data. It enables the separation of dependence structures from the marginal distributions. They provide a way to construct joint distributions with arbitrary margins and to impose a wide variety of dependence structures. They are fast gaining in popularity in engineering, financial applications, and recently also in the construction of synthetic populations (Jeong 2016 p. 6; Ye and Wang, 2018). The main challenge is to write a joint distribution of two (or more) random variables as the product of univariate margins and a distribution that describes the dependence structure assuming uniform margins. The merging of two distributions to approximate a true joint distribution is known as *coupling*. They are also applied to incorporate expert knowledge in joint distributions (O'Hagan, 2019). The coupling that maximizes the mutual information in two distributions minimizes the entropy of the joint distribution. It is therefore called a minimum-entropy coupling (Cicalese et al., 2019). It is the joint distribution that is farthest from independence between the random variables. In forecasting, empirical dependence structures may be adopted from historical records, known as nonparametric copula. Most copula models are developed for continuous distributions. Copula models for discrete probability distributions are in their infancy. A detailed discussion of copula models is beyond the scope of this paper.

2.4.2 Log-linear models

In the social sciences, a common approach to study dependencies in cross-classified count data (contingency tables) is to model the dependencies in the data (Bishop et al., 1975; Agresti, 2013, and Dobra and Mohammadi, 2018 for a Bayesian approach to log-linear models). The model separates the effects of marginal totals (main effects) from the association between the cross-classified variables (interaction effects). That model is the log-linear model. Applied to migration, the log-linear model decomposes a migration flow n_{ij} into an effect of the overall level of migration in the system of regions, an effect of the size of the population of i (origin), an effect of

the size of the population of j (destination), and an effect of the dependence or interaction between origin and destination. Since the model includes all lower-order terms contained in a higher-order term in the model, the total number of effects (and parameters) may exceed the number of cells in the contingency table, hence some parameters are redundant. To ensure that the parameters have unique values, restrictions are introduced, called identification or normalization restrictions. Each set of restrictions implies a particular coding scheme. In *treatment coding* (or dummy coding) one category is designed as the reference level, and the effects of other categories are measured relative to the effect of the reference category (first-order difference). In deviation coding or effect coding, each category gets its own parameter, and each parameter measures the effect relative the mean. In this coding scheme, the sum of the parameters is zero. For a discussion of coding schemes in the study of international migration, see Raymer (2007). A log-linear model that includes all lower-order terms is called a hierarchical log-linear model (Agresti, 2013). A reason for including lower-order terms is to prevent that the coding of the variables affect the statistical significance and interpretation of higher-order terms. Note that the values of the terms depend on the coding schemes adopted.

The model specifies a log-linear relation between cell counts and effects:

$$\ln n_{ij} = u + u_i^O + u_j^D + u_{ij}^{OD} \tag{2.45}$$

where O denotes origin and D destination. The parameter u measures the effect of the level of migration in the system (overall effect), u_i^O is the effect of level of emigration from region i (row effect), u_j^D the effect of level of immigration in j (column effect) and u_{ij}^{OD} is the interaction effect. The log-linear model is called *saturated* because it includes as many independent parameters as there are cells in the migration flow matrix.

Irrespective of the coding, first-order differences between parameters are unique. The odds is also unique:

$$\begin{aligned} \ln \frac{n_{ij}}{n_{1j}} = \ln n_{ij} - \ln n_{1j} &= u + u_i^O + u_j^D + u_{ij}^{OD} - (u + u_1^O + u_j^D + u_{1j}^{OD}) \\ &= u_i^O - u_1^O + u_{ij}^{OD} - u_{1j}^{OD} \end{aligned} \tag{2.46}$$

To see the relation between the log-linear model and (2.38), rewrite (2.38) as

$$exp\left[u^* + u_i^{*O} + u_j^{*D} + u_{ij}^{*OD}\right] = K a_i b_j exp\left[u^0 + u_i^{0O} + u_j^{0D} + u_{ij}^{0OD}\right]$$

with K a scaling factor that depends on the scaling of a_i (all i) and b_j (all j). The property of invariance of association implies that $u_{ij}^{*OD} = u_{ij}^{0OD}$, and

$$exp\left[u^* + u_i^{*O} + u_j^{*D}\right] = a_i b_j exp\left[u^0 + u_i^{0O} + u_j^{0D}\right]$$

If $K = \exp[u^* - u^0]$, then $u_i^{*O} = a_i \, exp\left[u_i^{0O}\right]$ and $u_j^{*D} = b_j \, exp\left[u_j^{0D}\right]$. It relates the parameters of the log-linear model of the matrix of migration estimates to the log-linear model parameters of the location preference matrix.

3 The multiregional model with preferences and restrictions: the macrosystem

The multiregional model is a prototype model of international migration inspired by the Schelling model of residential mobility (Schelling, 1971, 2006). Individuals occupy a place and they have agency, i.e. they act on their preferences. Location preferences are derived from subjective place

utilities. Individuals who are dissatisfied with their current place of residence desire to move. That desire motivates action, but when individuals attempt to turn the desire into action, they face several obstacles, e.g. the immigration restrictions imposed by governments. To incorporate the restrictions imposed on aggregate migration flows into the model, a distinction is made between a proposed migration and the actual migration. Individuals first apply for admission in their preferred place of residence. If their application is successful, i.e. they are admitted, they migrate. If they are not admitted, they adjust their location preferences and apply for admission next time. Some countries fill their quota during the first round, while other countries never reach their quota because they do not impose immigration quota or they are not sufficiently attractive to potential immigrants.

The main features of the Schelling model and its adaptation to international migration are presented in Section 3.1. The distinction between macrosystem and microsystem proves to be useful. Section 3.2 describes the model of the macrosystem. It is based on expected values. In Section 3.3 the model is applied to the global system of six regions. The microsystem is covered in Section 4.

3.1 The Schelling model and the adaptation to international migration

Schelling divides space in a rectangular grid system (checkerboard). A cell of the grid system or square of the checkerboard is vacant or occupied. Each cell is occupied by at most one individual. A cell may be interpreted as an address. The number of individuals is fixed and equal to the number of occupied cells. Individuals have personal attributes. In the original Schelling model, the attribute is race or ethnicity. Individuals have location preferences (residential preferences). The attractiveness of a location to an individual (and the place utility attached to a location) is determined by the population (ethnic) composition of the neighbourhood in which the location is situated. Schelling postulates that individuals prefer neighbourhoods in which they are not a minority. That principle is based on the theory that people prefer interaction with people who are similar (homophily). It is quantified by a preference level or tolerance level, i.e. the share of similar people should be at least 50 percent. Extensions of the model considered additional personal attributes, other neighbourhood characteristics, other tolerance levels and individual differences in tolerance level (population heterogeneity). Preferences are stated preferences, i.e. preferences individuals have but may not be realistic due to constraints.

Individuals have agency. They act on their preferences. An individual in a neighbourhood that does not meet the preference criterion (population composition) is not satisfied and intends to move to another neighbourhood. The individual moves to a vacant location. In the original model, a vacant location is selected at random from all vacant locations. Extensions introduced individual capabilities to collect information about the population composition of neighbourhoods and the selection of a location in a neighbourhood that is acceptable. Information acquisition results in an informed choice. A neighbourhood that is acceptable when the relocation decision is made, may no longer be acceptable after the move is made due to the relocation behaviour of other individuals in the system. Further extensions reduce and even remove vacancies, but gave individuals the capacity to collect information on the preferences of other individuals (e.g. through intermediaries such as real estate agents) and exchange or swap locations (home swapping).

Schelling's aim was to show that modest individual preferences coupled with a capability to act on the preferences can have far-reaching consequences at the population level. He showed that location preferences can lead to segregation. The model is also used to identify the conditions under which segregation occurs and to predict potential tipping points, when the population

composition in a neighbourhood has reached a point at which the majority population leaves and segregation is complete. The Schelling model triggered a great deal of theoretical, methodological and applied research. A number of authors approach the relocations in a grid system as a random walk and the segregation as the asymptotic behaviour of the random walk (Shin and Sayama, 2014). Several scholars formulated the Schelling model at a higher level of abstraction and exploited the analogy between the Schelling model and the Ising model in statistical mechanics (magnetism). In both models state changes (actions) are dependent upon attributes of adjacent cells (see e.g. Mantzaris et al., 2018).

The basic philosophy of the Schelling model remains valid when neighbourhoods are replaced by countries and groups of countries. The spatial structure is represented by a set of countries rather than a grid system. Individuals assign attributes to countries and attach place utilities. The place utilities depend on attributes of countries and not attributes of neighbouring countries. As in the Schelling model, individuals have no control over the attributes of countries and the policies enacted by national governments. Immigration quota are analogous to vacancies in the Schelling model. They are capacity constraints that restrict the freedom of movement A fundamental feature of immigration quota or caps on the number of immigrants admitted during a given period, is that they impose conditions that aggregate flows must satisfy. Since migrants have no control over admission policies and their implementations, they comply with the restrictions imposed or they find another strategy to reach their goals. Some circumvent the restrictions by crossing borders clandestinely or by overstaying authorized lengths of stay. It usually requires the support from a social network (Simon et al., 2018). Capacity constraints have been used before in models of international migration. Napierala et al. (2021) consider capacity constraints in forecasting asylum-related migration flows.

Individuals who intend to migrate respond to the presence of admission rules by applying for admission (entry visa). Individuals admitted migrate. Note that migration is no longer an event, but an outcome of a process consisting of stages: (a) migration intention, (b) application for admission, and (c) migration. Other stages may be added. The outcome of that process depends on an individual's intention, an admission or acceptance criterion, and on the application being accepted, which may partly be determined by chance. If the distribution of location preferences in a population is known, and the acceptance criterion is the availability of vacancies (quota not reached), then the procedure is quite simple and the solution can be obtained analytically. The method is presented in Section 3.2. If additional criteria are imposed, an analytical solution is no longer feasible and the solution should be obtained by simulation. For instance, if the requirement is to obtain a migration flow that satisfies marginal totals, reflects individual location preferences as accurately as possible, **and** is more probable than any other migration flow that satisfies these conditions, then simulation is the only option. The algorithm that operationalizes this procedure that meets these requirements resembles the Metropolis algorithm (Metropolis et al., 1953). The algorithm is widely used across the sciences. Its extension, the Metropolis-Hastings algorithm, represents the core of Markov chain Monte Carlo (MCMC), which is an algorithm for sampling from an unknown probability distribution with a probability density (continuous random variable) or probability mass function (discrete random variable) that is proportional to a known function.

Individuals share a location (region) and have location preferences. Unlike in the Schelling model, the factors that determine the perceived place utility are not identified. They are implicit in location preferences revealed by past migration flows. Data show that most people stay in their

current region of residence. They are satisfied, lack the resources to migrate, or are attached to their current place of residence for other reasons. The World Gallup Poll revealed that about 20 percent of the world population desires to emigrate. Stated preferences do not fully consider the different types of restrictions that apply, however. Due to different types of restrictions, a little over one percent plans to emigrate and less than one percent prepare for emigration. Globally, not more than 1.2 percent of the population live in a country other than their country of residence 5 years ago (Abel and Cohen, 2019) and 0.6 percent live in another region than their region of residence 5 years ago (Tables B.9 and B.10). If changes in location preferences are allowed, then immigration restrictions may affect all location preferences, including the preference to stay in the current region of residence. In other words, immigration restrictions affect the entire probability distribution of location preferences. As shown in Section 2 of the paper, some key features of the distribution remain unchanged.

An individual with a preferred region of residence other than the current region of residence, proposes a migration to the preferred region of residence. The proposed destination is selected at random from possible destinations or the proposal is an informed choice. In case of random selection, the destination is obtained by simple multinomial sampling of possible destinations using the probability mass function of the location preferences. The proposal destination is accepted if it meets the conditions imposed by the immigration restrictions. For instance, the proposal may be accepted of the immigration quota is not reached yet or if the immigrants has certain desirable attributes. An important feature of the model is that the acceptance of an individual's proposal depends on proposals and actions of other individuals in the system of regions. If an individual is not admitted, then the individual stays in the region of residence and gets another chance to propose a destination in the next step of the simulation, provided some immigration countries have vacancies or a more liberal admission policy. Once a proposal is accepted, the individual relocates to the selected region of destination.

Extensions may give individuals the capability to acquire information about admission policies and adapt their location preferences based on that information. Individuals may also interact with other individuals in the current region of residence and in the possible regions of destination (diaspora). They may exchange information, opinions, resources, goods and services with members of a social network. Relationships between individuals are not considered in this paper, but the microsimulation is sufficient flexible to introduce relationships and social network support. Extensions of the model may also consider more complex and realistic decision processes that involve cognitive and social processes. Finally, extensions may consider location swapping. When individuals exchange locations, the marginal totals of the migration flow matrix are not affected. The origin-destination interaction is affected, however.

3.2 Method

The multiregional model with preferences and restrictions is visualized in a flow diagram (Figure 3.1). Two types of actors at two different levels of aggregation are distinguished. At the micro-level, individuals are the actors. At the country level, governments are the actors.

Let ***Pop***(t) be the column vector of number of individuals by country (region) of residence at time t and ***Pref*** the location preference matrix. All vectors are column vectors unless specified differently. An element $Pref_{ij}$ of ***Pref*** is the proportion of the residents of country i that denotes country j as the preferred country of residence. The diagonal elements denote the proportions satisfied with

their country of residence. The matrix ***Pref0***, introduced later, is ***Pref*** with the diagonal elements set to zero. It is the location preference matrix excluding the desire to stay. The number of individuals by country of residence at t and preferred country of residence is the matrix $\boldsymbol{SF}(t)$:

$$\boldsymbol{SF}(t) = diag[\boldsymbol{Pop}(t)]\ \boldsymbol{Pref} \tag{3.1}$$

An element $SF_{ij}(t)$ of $\boldsymbol{SF}(t)$ denotes the number if country in i at time t that prefers to live in country j. The diagonal elements of $\boldsymbol{SF}(t)$ show the number of individuals in their preferred country (or region) of residence. They are satisfied. The off-diagonal elements show the number of dissatisfied individuals, by country of residence at t and preferred country of residence. The row sum of the off-diagonal elements is the number of dissatisfied individuals in country i at t. The column sum is the total number of dissatisfied individuals interested in moving to country j. If all these people apply for admission, it is the number of applications received by country j. $\boldsymbol{SF}(t)$ with the diagonal elements set to zero is denoted by $\boldsymbol{SF}0(t)$. The matrix gives the number of *dissatisfied* people by current country of residence and preferred country of residence.

Destination countries have different policies and apply different criteria to admit people. Suppose they impose immigration quota but they are indifferent about who applies for a residence permit and who they admit. This means that immigrants are selected randomly. The probability that country j accepts the application of a resident of country i to fill the immigration quota depends on the number of individuals in i with j as the preferred country of residence (j≠i). It residents of i submit many more applications to country j then residents of other countries, then their share in the admissions is higher than that of other countries. The proportions admitted during the unit interval following time t are obtained by dividing the elements of $\boldsymbol{SF}0(t)$ by their column sums:

$$\boldsymbol{Padm}(t) = \boldsymbol{SF}0(t)\left[diag[\mathbf{1}'\ \boldsymbol{SF}0(t)]\right]^{-1} \tag{3.2}$$

An element $Padm_{ij}(t)$ of the admission matrix is an estimate of the probability that an individual admitted in country j originates from country i. The admission matrix is also referred to as recruitment matrix and the admission probability as a recruitment probability. Note that the admission probabilities depend on the number of people in each country and their location preferences. If the individuals admitted are selected randomly from the applications, then the relation between recruitment probabilities and individual location preferences is

$$\boldsymbol{Padm}(t) = diag[\boldsymbol{Pop}(t)]\ \boldsymbol{Pref0}\ \ diag\{[\mathbf{1}'\ diag[\boldsymbol{Pop}(t)]\ \boldsymbol{Pref0}]^{-1}\}$$

The total number of people admitted in j or residence permits issued by j is determined by the immigration quota. If the number of applications country j receives exceeds the immigration quota, the number of applications honoured and people selected for immigration is the immigration quota, while the remaining applications are rejected. If the immigration quota of j is higher than the number of people interested in j, all applicants are admitted and j and an unfilled quota remains. The number of applicants selected, by country of destination or immigration, is

$$\boldsymbol{nadm}(\mathrm{t}) = \min[\boldsymbol{quota}(t), [\mathbf{1}'\ \boldsymbol{SF}0(t)]'] \tag{3.3}$$

where $\boldsymbol{quota}(t)$ is the column vector of immigration quota by country and $\mathbf{1}'$ is a row vector of ones. Note that $\mathbf{1}'\ \boldsymbol{SF}0(t)$is a row vector and $\left[\mathbf{1}'\ \boldsymbol{SF}0(t)\right]'$ is a column vector. The minimum values are selected elementwise. For countries without immigration quota, the quota is fixed at a very

high number so that the number of applications will never exceed the quota. The number of applications by country of origin and country of admission is

$$\boldsymbol{nadmOD}(t) = \boldsymbol{Padm}(t)\ diag[\boldsymbol{nadm}(t)] \tag{3.4}$$

The column sums of the matrix $\boldsymbol{nadmOD}(t)$ gives the number of individuals each country admits and the row sums gives the number of individuals by country of current residence admitted in their preferred country of residence. The column sums are $\boldsymbol{nadm}^{des}(\mathrm{t}) = \mathbf{1}'\ \boldsymbol{nadmOD}(t)$. The row sums are $\boldsymbol{nadm}^{orig}(t) = \boldsymbol{nadmOD}(t)\ \mathbf{1}$ with $\mathbf{1}$ a column vector of ones and $\mathbf{1}'$ a row vector of ones.

The number of individuals who are dissatisfied with their country of residence and apply for immigration in their preferred country of residence, but whose application is rejected, is

$$\boldsymbol{notadmOD}(t) = \boldsymbol{SF}0(t) - \ \boldsymbol{nadmOD}(t) \tag{3.5}$$

The number of individuals not admitted, by immigration country, is $\boldsymbol{notadm}^{des}(t) = \mathbf{1}'\ \boldsymbol{notadmOD}(t)$. The number not admitted to their preferred country of residence, by current country of residence is $\boldsymbol{notadm}^{orig} = \boldsymbol{notadmOD}(t)\ \mathbf{1}$.

An admission is followed by a migration. Individuals admitted to their preferred country of residence migrate during the time interval from t to t+1. Hence $\boldsymbol{nadmOD}(t)$ is also a migration matrix. It gives the number of migrations from country of residence at t to the preferred country and consequently the country of residence at t+1. Individuals whose application is not honoured remain, at least temporarily, in their current country of residence and continue to be dissatisfied. The model captures the phenomenon, revealed by several empirical studies, that insufficient legal pathways for immigration increases the level of dissatisfaction among people with a migration intention. They may apply for admission again later. Repeated rejections build frustrations, which may result in a decision to stay or irregular migration. In the model, irregular migration is a consequence of dissatisfaction due to the inadequacy of legal pathways for immigration, a link revealed by empirical evidence (e.g. Czaika and Hobolth, 2016; Clemens and Gough, 2018; Barslund et al., 2019).

At the end of this round of applications and admissions, some countries have reached their immigration quota, while other countries received less immigrants than the quota allowed. The unfilled quota is $\boldsymbol{quota}'(t) - \mathbf{1}'\boldsymbol{nadmOD}$.

The population of country j at t+1 consists of (a) residents of j who were satisfied with their place of residence and stayed in the country (the great majority), (b) residents of countries other than j who were dissatisfied with their place of residence, preferred to move to j, applied for admission, whose application was accepted, and who subsequently moved to j, and (c) residents of j who were dissatisfied with j, applied for admission in their preferred country of residence, but were not successful. Hence:

$$\boldsymbol{nPop}(t+1) = diag[\boldsymbol{SF}(t)] + \ \boldsymbol{nadm}(\mathrm{t}) + \ \boldsymbol{notadm}^{orig}(t) \tag{3.6}$$

with $diag[\boldsymbol{SF}(t)]$ the diagonal of the matrix $\boldsymbol{SF}(t)$[9].

The country of residence of the population at t+1, by country of residence at t, is

$$\boldsymbol{nloc}(t, t+1) = diag(diag[\boldsymbol{SF}(t)]) + \boldsymbol{nadmOD}(t) + \boldsymbol{notadm}^{orig} \tag{3.7}$$

Suppose individuals whose applications are rejected may apply to a different country during the same period. Location preferences are updated accounting for the countries that reached their quota and are no longer possible destinations. To distinguish between countries with filled quota and those with unfilled quota, an indicator variable is introduced. Let $quotec_i$ denote the value of the indicator for country i. It is 1 if i reached its quota and 0 otherwise. The vector of indicator variables is $\boldsymbol{quotac}$. If individuals adhere to their original location preferences but adjust the choice set, then the matrix of updated location preferences is

$$\boldsymbol{Prefu} = \left[diag[\boldsymbol{Pref}\ diag(\boldsymbol{quotac})\ \mathbf{1}]\right]^{-1} \boldsymbol{Pref}\ diag(\boldsymbol{quotac}) \tag{3.8}$$

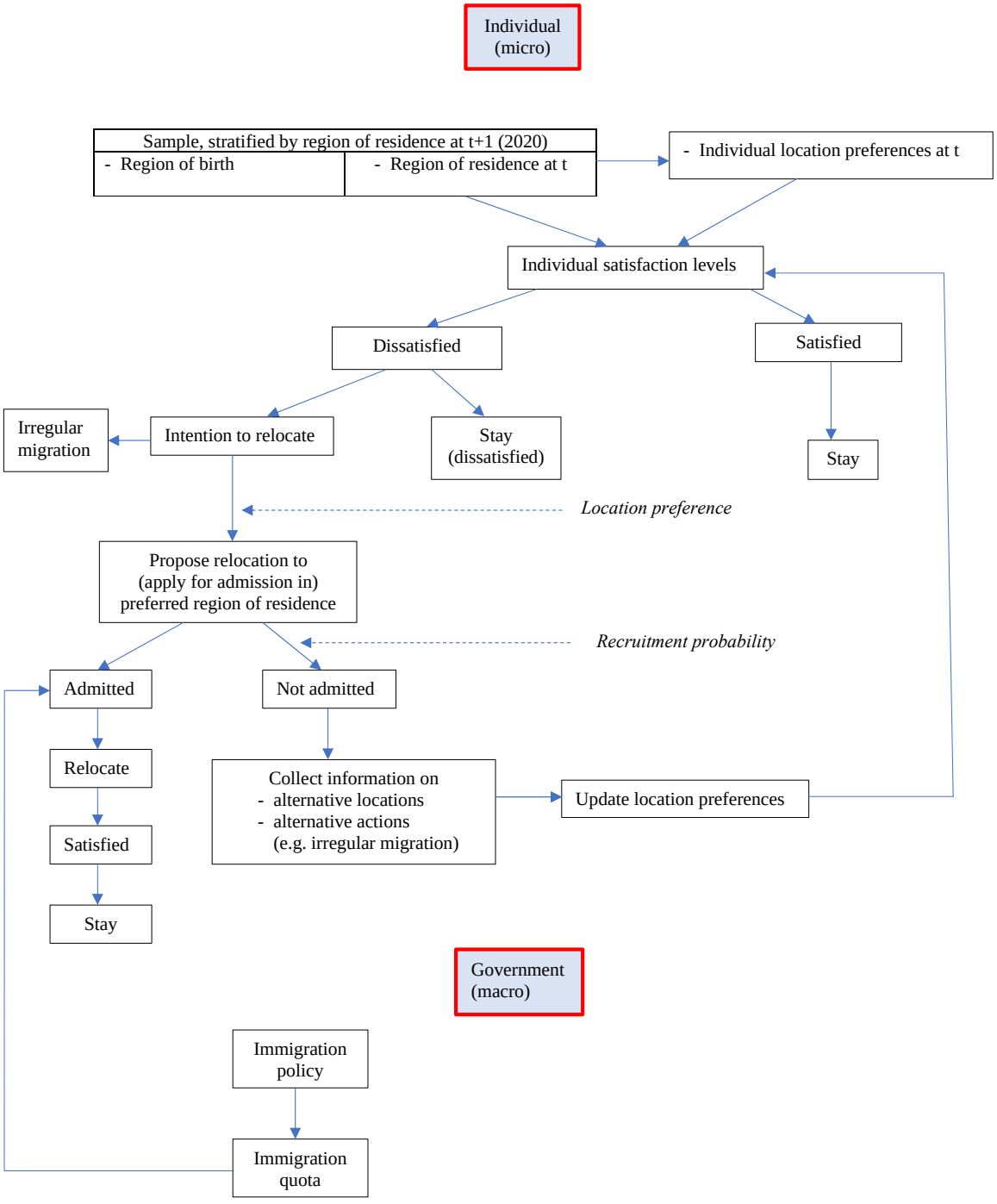
The first term on the right-hand-side is the inverse of a diagonal matrix. The diagonal consists of row sums of the updated preference matrix. If individuals may update their location preferences multiple times during a period, then the above function is applied recursively.

Adherence to the initial location preferences implies that nearly all individuals who experience a rejection will decide to remain in their original country of residence. To remove that option, ***Pref*** is replaced by ***Pref0*** in the computation of ***Prefu***.

To assess the quality of the flow estimates $\boldsymbol{nloc}(t, t+1)$, the entropy is used as a measure of distance between the estimated flow and the reported flow for the period 2015-20. Entropy is a global measure of diversity and uncertainty. It is a popular measure in the study of complex phenomena. Its definition is given by equation (2.19). High entropy implies high uncertainty. Knowledge reduces uncertainty. In this paper, the knowledge consists of marginal totals and location preferences. Entropy maximization produces most probable flow estimates given the available knowledge. They leave as much uncertainty as possible while accounting for the prior knowledge. In the following subsection, it will be shown how the stepwise addition of information on migration flows reduces the uncertainty or, equivalently, increases the expected information content of the estimated migration flow.

[9] If ***A*** is a matrix, then ***diag(A)*** denotes the diagonal of ***A***. If ***x*** is a vector, the ***diag(x)*** denotes a diagonal matrix with ***x*** in the diagonal.

Figure 3.1


Individual (micro)
Sample, stratified by region of residence at t+1 (2020)
- Region of birth
- Region of residence at t
- Individual location preferences at t
Individual satisfaction levels
Dissatisfied
Satisfied
Irregular migration
Intention to relocate
Stay (dissatisfied)
Stay
Location preference
Propose relocation to (apply for admission in) preferred region of residence
Recruitment probability
Admitted
Not admitted
Relocate
Satisfied
Stay
Collect information on
- alternative locations
- alternative actions
(e.g. irregular migration)
Update location preferences
Government (macro)
Immigration policy
Immigration quota

3.3Application

Picture the world as a system of six regions and assume that regions impose a cap on the number of immigrants during a given period. Assume that the immigration quota imposed is based on the immigrations during a base period, which is 2015-2020. The assumption is unrealistic and is made for illustrative purposes only. A more realistic assumption is that some regions do not impose immigration quota. In that case, the immigration quota is set at a very large number. Table B.9 shows the number of residents in each region at the start and the end of the base period (mid-2015 and mid-2020). The total, 7.8 billion, is the population in 2020 by region of residence in 2020 and region in 2015. In the Abel-Cohen estimates, births and deaths during the period are added to and subtracted from the diagonal elements and return migration is accounted for. The row totals represent the population in 2015 and the column totals the population in 2020 by region of residence. The same row totals are shown in Table 3.1.

Location preferences are assumed to be the preferences revealed by the migration flow matrix in the reference period 1995-2000. The location preferences are shown in Table B.12. In the absence of immigration quota, the location preferences determine where the population of 2015 (and present in 2020) lives in 2020: $\boldsymbol{Pop}^a(t+1) = [\boldsymbol{Pref}]'\boldsymbol{Pop}(t)$. $\boldsymbol{Pop}^a(t+1)$ differs from the observed population in 2020 ($\boldsymbol{Pop}(t+1)$) because the revealed location preferences in 1995-2000 differ from the migration flow in 2015-19. The population by region of residence at t and preferred region of residence is given by equation (3.1) and shown in Table 3.1:

Table 3.1 Population by region of residence and preferred region of residence

	Preferred region of residence in 2020						
Region 2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	515.047	2.160	0.521	0.796	1.543	1.500	521.567
USCan	0.963	358.055	2.079	0.115	1.366	0.183	362.761
LatAm	0.843	7.749	647.368	0.008	0.176	0.031	656.174
Africa	3.590	1.190	0.037	1336.722	1.176	0.194	1342.908
Asia	3.307	5.113	0.166	0.490	4610.953	4.677	4624.705
Rest	2.163	0.472	0.030	0.032	1.608	257.427	261.733
Sum	525.913	374.739	650.200	1338.162	4616.822	264.013	7769.849

The diagonal shows the number of people who are satisfied with their region of residence in 2015. The off-diagonal elements show the number of people dissatisfied with their region of residence, by region of residence in 2015 and preferred region of residence.

Assume that all regions impose immigration quota and that the quota a region imposes is equal to its population in 2020 minus the number of their residents in 2015 who prefer to stay in the region, based on the location preferences revealed during the reference period 1995-2000. During the reference period, 98.75 percent of the residents present in EU+ at the beginning of that period (1995) are in EU+ at the end of the period (2000). If that location preference applies to the period 2015-19, then 515.05 million of the 521.57 million residents in 2015 prefer to stay in EU+. The figure is somewhat higher than the 514.32 million estimated by Abel and Cohen (2019) (Table B.9). The difference is small, however, indicating stable revealed preferences. The immigration quota is the difference between the EU+ population in 2020 (527.55 million; Table B.9) and the number of people with a preference to stay in EU+ (514.32 million). The quota obtained this way is 12.51 million, higher than the 10.87 million (Table 3.1) people in 2015 in other regions of the world who are dissatisfied with their place of residence and prefer to relocate to EU+. The reason for the difference is that the number of stayers estimated by Abel and Cohen (514.32 million) is less than

the number of stayers expected by the revealed location preferences. In other words, residents of EU+ in 2015 are a little less in favour of the EU+ than residents in 1995.

The picture differs for USCan. For the period 2015-19, Abel and Cohen estimate that, of the population of 362.76 million in 2015, 355.59 million are also in USCan in 2020. Revealed location preferences give 358.06 million, however. Residents of USCan in 2015 have a higher preference for staying in USCan than residents in 1995. The higher preference for staying is probably due to the increased immigration restrictions during that period. It is well-known that immigration restrictions have the unintended side effect that immigrants are less likely to leave the country because of the lower prospect to return (Massey et al., 2014; de Haas et al., 2019). The computed immigration quota for USCan is 10.69 million, the difference between the population in 2020 (368.75 million) and the estimated number of stayers based on the revealed preference (358.06 million; Table 3.1). The number is considerably lower than the number of people in the rest of the world who are dissatisfied with their location and prefer USCan (16.68 million; Table 81). A consequence of that disparity is a rejection by USCan of 6 million applications for admission (one third).

A total of 3.59 million persons in Africa and 0.84 million in Latin America and the Caribbean prefer EU+, while respectively 1.19 million and 7.75 prefer USCan, respectively.

Table 3.2 summarizes the results obtained until now. The first columns show the population in 2015, present in 2020, and the population in 2020 present in 2015. The number satisfied in 2015, by region of residence, is shown in column 3. The next two columns show the number dissatisfied, first by region of residence in 2015 (dissat015) and next by region of preference (dissatD15). Note that globally the proportion of the population that is dissatisfied with their current region of residence is 0.6 percent (44.28/7769.85). The immigration quota are shown next. The number selected, by the destination region, (nselectedD) is the number dissatisfied (dissatD15) or the quota, whatever is lowest. It is computed using equation (3.3). The next columns show the number of individuals selected and not selected, by preferred region of residence. The last column indicates the size of the unfilled quota. Globally, 83 percent of the applications for admission are accepted. Three regions do not reach their immigration quota as computed in this section: EU+, Latin America and the Caribbean, and Africa.

Table 3.2 Summary of results

	pop2015	pop2020	satisfied15	dissatO15	dissatD15	quota	selectedD	notselectedD	unfilledquota
EU+	521.57	527.55	515.05	6.52	10.87	12.51	10.87	0.00	1.64
USCan	362.76	368.74	358.05	4.71	16.68	10.69	10.69	5.99	0.00
LatAm	656.17	653.56	647.37	8.81	2.83	6.19	2.83	0.00	3.36
Africa	1342.91	1340.59	1336.72	6.19	1.44	3.87	1.44	0.00	2.43
Asia	4624.71	4616.03	4610.95	13.75	5.87	5.08	5.08	0.79	0.00
Rest	261.73	263.37	257.43	4.31	6.59	5.94	5.94	0.65	0.00
Sum	7769.85	7769.85	7725.57	44.28	44.28	44.28	36.85	7.43	7.43

The regions of origin of applicants admitted to a region are determined by the total number admitted and the recruitment matrix. The matrix of recruitment probabilities, computed using equation (3.2), is shown in Table 3.3. One third of the individuals admitted in EU+ during the reference period (1995-2000) is from Africa, 30 percent from Asia and 20 percent from the rest of the world.

Table 3.3 Admission probabilities

		Region admitting applicants					
Region 2015		EU+	USCan	LatAm	Africa	Asia	Rest
	EU+	0.00000	0.12947	0.18396	0.55248	0.26286	0.22781
	USCan	0.08867	0.00000	0.73394	0.07953	0.23279	0.02782
	LatAm	0.07758	0.46445	0.00000	0.00562	0.02998	0.00472
	Africa	0.33035	0.07131	0.01295	0.00000	0.20032	0.02953
	Asia	0.30434	0.30647	0.05850	0.33996	0.00000	0.71012
	Rest	0.19907	0.02830	0.01065	0.02242	0.27405	0.00000
	Sum	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Table 3.4 shows, for each region of origin, how many dissatisfied individuals are admitted to their preferred region of residence. The figures are obtained by equation (3.4). The table also shows the number not admitted and the total. The row sum gives the number of applications granted and rejected by region of origin. The picture for the EU+ is very different from that of USCan. In USCan, almost half of those admitted are from Latin America and the Caribbean. The proportion originating from Asia is the same as in EU+.

Table 3.4 Number of individuals whose application for admission are accepted/rejected, by region of origin and preferred region of residence (in million)

A. Application for admission granted

	PreferredRegion						
CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0.00	1.38	0.52	0.80	1.33	1.35	5.39
USCan	0.96	0.00	2.08	0.11	1.18	0.17	4.50
LatAm	0.84	4.96	0.00	0.01	0.15	0.03	6.00
Africa	3.59	0.76	0.04	0.00	1.02	0.18	5.58
Asia	3.31	3.28	0.17	0.49	0.00	4.22	11.46
Rest	2.16	0.30	0.03	0.03	1.39	0.00	3.92
Sum	10.87	10.69	2.83	1.44	5.08	5.94	36.85

B. Application for admission rejected

	PreferredRegion						
CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0	0.78	0	0	0.21	0.15	1.13
USCan	0	0.00	0	0	0.18	0.02	0.20
LatAm	0	2.78	0	0	0.02	0.00	2.81
Africa	0	0.43	0	0	0.16	0.02	0.61
Asia	0	1.84	0	0	0.00	0.46	2.30
Rest	0	0.17	0	0	0.22	0.00	0.39
Sum	0	5.99	0	0	0.79	0.65	7.43

C. Total applications for admission

	PreferredRegion						
CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0.00	2.16	0.52	0.80	1.54	1.50	6.52
USCan	0.96	0.00	2.08	0.11	1.37	0.18	4.71
LatAm	0.84	7.75	0.00	0.01	0.18	0.03	8.81
Africa	3.59	1.19	0.04	0.00	1.18	0.19	6.19
Asia	3.31	5.11	0.17	0.49	0.00	4.68	13.75
Rest	2.16	0.47	0.03	0.03	1.61	0.00	4.31
Sum	10.87	16.68	2.83	1.44	5.87	6.59	44.28

Individuals who are admitted in their preferred region of residence migrate to that region, while individuals whose application for admission is rejected stay in their initial region of residence (and remain dissatisfied with their place of residence). The distribution of the population by initial region of residence (2015), region of residence at the end of the procedure, and level of satisfaction at that time is shown in Table 3.5. Notice that 0.2 million individuals in USCan are dissatisfied. The reason is that they were dissatisfied with USCan in 2015 and preferred to relocate to Asia or the Rest of the World, but their application was rejected (see Table 3.4, panel B). In EU+, 1.13 million individuals are dissatisfied with their region of residence. Most prefer USCan (0.78 million), followed by Asia (0.21 million) and rest of the world (0.15 million). The number of applications rejected by USCan is the total number of applications (2.61 million; Table 3.1) and the number accepted (1.38 million; Table 3.4A). In Latin America and the Caribbean, almost all of the 2.81 million residents who applied unsuccessfully during the period 2015-19 to any other region, were rejected by USCan (2.78 million).

The population in 2020 predicted by the multiregional model is shown in the column totals in Table 3.5 C. The figures differ slightly from the actual population in 2020, which is shown in the column totals in Table 3.6. The difference is due to the number of dissatisfied individuals. A total of 0.2 million residents of USCan in 2015 were dissatisfied and applied for admission to another region, but were not admitted. They remain in USCan, which explains that the figure in Table 3.5 C (368.95) exceeds the observed figure (368.74) by 0.2 million. Notice that the observed number of satisfied residents in USCan in 2020 is 368.74 (Table 3.5 A, sum of second column). For EU+, the reason for the difference is more complex. It is the difference between the unfilled quota (1.64; Table 3.2) and the number of EU+ residents in 2015 that was dissatisfied and preferred to move to a region other than the EU+ (mainly USCan), but their application was rejected (1.13; Table 3.5 B). The difference is 0.51, which is also the difference between the population in EU+ in 2020 (527.55 million) and the value predicted by the multiregional model (527.04). If the number of dissatisfied residents of EU+ unable to move to their preferred region of residence would be equal to the unfilled immigration quota of EU+, then the model would produce a perfect prediction of the observed number of residents in 2020.

The population in 2020 predicted by the model tends to the observed population in that year (i.e. estimated by Abel and Cohen) if dissatisfied residents get the opportunity to update their location preferences and move to their updated preferred region of residence. The 0.2 million dissatisfied individuals in USCan would move to regions with unfilled immigration quota. It would eliminate dissatisfaction of USCan residents with their region of residence and would reduce the unfilled immigration quota in the regions that receive these immigrants. The capacity of individuals to update location preferences in light of information that some regions have reached their immigration quota eliminates the unfilled immigration quota and ensures that the population in 2020 predicted by the model is equal to the observed population.

Table 3.6 shows the population by region in 2015 and 2020, estimated by Abel and Cohen, augmented by stayers in their country of birth, derived from UN population data (see Annex B). A comparison of Table 3.6 and Table 3.5C reveals that fewer people stayed in EU+ than predicted by the model (514.31 versus 516.18). The difference is explained by a higher than expected migration of EU+ residents (2015) to Africa and Asia. It is very likely that these people are return migrants. The model also predicts more immigrants from Africa in EU+ than estimated by Abel and Cohen (3.59 million versus 2.71 million). The model, however, underestimates the number of immigrants from Asia (3.31 million versus 5.73 million). The model also underestimates the number of immigrants in the USCan originating from Asia and Africa, although it predicts very well the number of immigrants from Latin America and the Caribbean. The differences are caused by the

revealed location preferences used to predict migration flows and indicate a shift in location preferences. For instance, the data reveal a shift in revealed destination preferences of African residents in the last decades. The EU+ has become less popular, while USCan has become more popular. It is part of a diversification of African emigration, which, according to Flahaux and De Haas (2016) is partly driven by immigration restrictions by European states.

Table 3.5 Number of people by region of residence in 2015, region of residence in 2020, and level of satisfaction with their location in 2020 (in million)

A. Satisfied

	CurrentRegion						
InitialRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	515.05	1.38	0.52	0.80	1.33	1.35	520.44
USCan	0.96	358.05	2.08	0.11	1.18	0.17	362.56
LatAm	0.84	4.96	647.37	0.01	0.15	0.03	653.36
Africa	3.59	0.76	0.04	1336.72	1.02	0.18	1342.30
Asia	3.31	3.28	0.17	0.49	4610.95	4.22	4622.41
Rest	2.16	0.30	0.03	0.03	1.39	257.43	261.35
Sum	525.91	368.74	650.20	1338.16	4616.03	263.37	7762.42

B. Not satisfied

	CurrentRegion						
InitialRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	1.13	0.0	0.00	0.00	0.0	0.00	1.13
USCan	0.00	0.2	0.00	0.00	0.0	0.00	0.20
LatAm	0.00	0.0	2.81	0.00	0.0	0.00	2.81
Africa	0.00	0.0	0.00	0.61	0.0	0.00	0.61
Asia	0.00	0.0	0.00	0.00	2.3	0.00	2.30
Rest	0.00	0.0	0.00	0.00	0.0	0.39	0.39
Sum	1.13	0.2	2.81	0.61	2.3	0.39	7.43

C. Total

	CurrentRegion						
InitialRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	516.18	1.38	0.52	0.80	1.33	1.35	521.57
USCan	0.96	358.26	2.08	0.11	1.18	0.17	362.76
LatAm	0.84	4.96	650.18	0.01	0.15	0.03	656.17
Africa	3.59	0.76	0.04	1337.33	1.02	0.18	1342.91
Asia	3.31	3.28	0.17	0.49	4613.25	4.22	4624.71
Rest	2.16	0.30	0.03	0.03	1.39	257.81	261.73
Sum	527.04	368.95	653.01	1338.77	4618.33	263.75	7769.85

The results should also be compared to the most probable migration flow that satisfies the marginal totals of the reported migration flow 2015-19 (Table 3.6) and reflects as accurately as possible the location preferences revealed by the 1995-00 migration flow. The estimates are obtained by the method of entropy maximization (Section 2.3). Abel's R package migest is used. The estimates are shown in Table 3.7. An interesting relation exists between the original location preferences (Table C.5 in Annex C) (A), the location preferences revealed by the migration estimates (B), and the location preferences revealed by the relocation pattern of individuals who successfully applied for admission to their preferred region of residence (Table 3.5A) (C). First, compare A and B. The location (destination) preferences revealed by B differ from the original location preferences. The relative location preferences differ too. Ratios of relative preferences, i.e. odds ratios, do not differ, however. It is a consequence of the multiplicative invariance of association, discussed in Section 2. The estimation method (IPF) does not preserve the relative

location preferences, but preserves the ratios of relative preferences. In other words, first-order differences are not preserved, but second-order differences are. Next, compare A and C. The location preferences revealed by the relocation of successful applicants differ from the original location preferences. Some relative preferences are different and other are equal. The relative destination preferences revealed by the migration of individuals admitted in their preferred region of residence are equal to the original relative location preferences only for destination with unfilled immigration quota. Equality indicates that individuals could realize their location preferences because immigration quota do not represent binding constraint.

Table 3.6 Population by region in 2015 and 2020 (in million)

	Region in 2020						
Region 2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	514.31	1.28	0.90	1.33	2.02	1.72	521.57
USCan	1.38	355.59	2.92	0.33	2.10	0.43	362.76
LatAm	1.57	4.91	649.47	0.01	0.13	0.08	656.17
Africa	2.71	1.13	0.02	1337.94	0.98	0.13	1342.91
Asia	5.73	5.49	0.21	0.89	4608.90	3.47	4624.71
Rest	1.84	0.34	0.03	0.08	1.90	257.53	261.73
Sum	527.55	368.74	653.56	1340.59	4616.03	263.37	7769.85

Source: UN data and estimates by Abel and Cohen (2019)

Table 3.7 Population by region in 2015 and 2020 reflecting location preferences (in million)

	Region in 2020						
Region 2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	515.96	1.53	0.53	1.01	1.28	1.25	521.57
USCan	1.36	356.40	2.98	0.21	1.60	0.21	362.76
LatAm	0.83	5.39	649.77	0.01	0.14	0.03	656.17
Africa	2.83	0.66	0.03	1338.49	0.77	0.13	1342.91
Asia	3.98	4.35	0.20	0.75	4610.72	4.69	4624.71
Rest	2.59	0.40	0.04	0.05	1.60	257.05	261.73
Sum	527.55	368.74	653.56	1340.52	4616.11	263.37	7769.85

The data used to estimate the migration flow in 2015-19 contribute differently to the estimates. To determine the contributions of data sources, the extent to which they reduce the uncertainty is quantified. The entropy is used to quantify the uncertainty. The entropy of a directional migration flow (macrostate) is maximum when all individuals in a population have the same probability of relocation from j, irrespective of their current residence. In that case, all migrant transition probabilities are the same and the entropy of the 6-region systems is

$$H^{uniform} = -\sum_{i,j} p_{ij}^{uniform} \ln p_{ij}^{uniform}$$

with $p_{ij} = 1/36$, which is $H = ln\,(36) = 3.58$. Knowledge of the marginal distributions (population distribution in 2015 and population distribution in 2020) reduces the entropy of the macrostate to 2.52. Knowledge of the location preference matrix, revealed by the 1995-00 migration flow, reduces the uncertainty further to an entropy of 1.30. This value is computed as $H^{estim} = -\sum_{i,j} p_{ij}^{estim} \ln p_{ij}^{estim}$ with p_{ij}^{estim} computed by equation (2.38): $p_{ij}^{estim} = \frac{n_{ij}^{estim}}{n_{++}^{*}} = \frac{a_i b_j n_{ij}^{0}}{\sum_{ij} a_i b_j n_{ij}^{0}}$. The reduction from 2.52 to 1.30 is due to the information on the dependence between origin and destination contained in the preference matrix. Hence information on revealed location preferences substantially reduces the uncertainty in the estimates of migration flows. The entropy of the

reported migration flow during the period 2015-19 (1.3016) is close to the entropy of the migration flow during the period 1995-00 and used as revealed preferences (1.3085) and the emigration flow estimates for the period 2015-19 (1.2975).

The precise contribution of information to the reduction of uncertainty in migrations flows is quantified by the information gain or Kullback-Leibler information divergence. The KL information divergence between the reported migration flow in 2015-19 and the absence of any knowledge (uniform distribution) on that migration flow is:

$$D_{KL}(p_{ij}^{15-19} \Big\| p_{ij}^{uniform}) = \sum\nolimits_{ij} p_{ij}^{15-19}\, ln \frac{p_{ij}^{15-19}}{p_{ij}^{uniform}} = 2.2819$$

It is equal to the weighted sum of the uniformly distributed relocation probabilities $p_{ij}^{uniform}$ with weights the relocation probabilities in 2015-19 minus the entropy of the reported flow in 2015-19:

$$D_{KL}(p_{ij}^{15-19} \Big\| p_{ij}^{uniform}) = \sum\nolimits_{ij} p_{ij}^{15-19}\, ln \frac{1}{p_{ij}^{uniform}} - \left[\sum\nolimits_{ij} p_{ij}^{15-19} \ln \frac{1}{p_{ij}^{15-19}}\right] = 3.5835 - 1.3016$$
$$= 2.2819$$

$D_{KL}(p_{ij}^{15-19} \Big\| p_{ij}^{indep}) = \sum_{ij} p_{ij}^{15-19}\, ln \frac{p_{ij}^{15-19}}{p_{ij}^{indep}} = 1.2207$. Knowledge of the marginal distributions of the 2015-19 flow reduces the uncertainty from 2.28 to 1.22. Knowledge of the joint distribution of region of residence in 1995 and region of residence in 2000 reduces the uncertainty further to 0.0083, a significant reduction. It is the KL information divergence $D_{KL}(p_{ij}^{15-19} \Big\| p_{ij}^{95-00}) = \sum_{ij} p_{ij}^{15-19}\, ln \frac{p_{ij}^{15-19}}{p_{ij}^{95-00}} = 0.0083$. That major reduction is possible only if the dependence between origin and destination in 2015-19 is similar to that in 1995-00. Notice that the joint distribution is different from the location preferences used in the previous analysis, which are conditional probabilities (location preferences conditional on region of residence at the start of the interval). The joint distribution carries more information than the conditional distribution. Knowledge of the conditional distribution would reduce the uncertainty from 1.22 to 0.53: $D_{KL}(p_{ij}^{15-19} \Big\| p_{ij}^{locpref}) = \sum_{ij} p_{ij}^{15-19}\, ln \frac{p_{ij}^{15-19}}{p_{ij}^{locpref}} = 0.5326$. Note that the joint distributions of region of residence in 1995 and region of residence in 2000 can be obtained by multiplying the conditional distributions for residents of region i in 1995 (i=1, 2, …, r) by the probabilities that a randomly selected member of the population in 1990 resides in i, which is the state probability introduced in Section 2. Hence the uncertainty reduction from 0.5326 to 0.0083 can be attributed to the knowledge of the spatial population distribution in 1995. A further reduction of uncertainty about the migration flow in 2015-19

is possible if one knows the most probable migration flow in 2015-19 that is consistent with the reported population in 2015 and the population in 2020 (constraints) and best reflects the location preferences revealed by the 1995-00 migration flow matrix. The estimates are obtained by KL information divergence minimization (equation 2.36). The uncertainty about the 2015-19 migration flow remaining in the presence of the most probable estimates that satisfy the conditions imposed above is 0.0002 (more precisely, 0.000249). It is considerably lower than 0.0083. The further reduction can be attributed to the addition of the marginal totals of the 2015-19 migration flow. The estimation procedure (entropy maximization) is needed to merge the dependence structure in the 1995-00 migration matrix and the marginal totals of the 2015-19 migration matrix. A further reduction to 0 can only be realized by complete information on the migration flow in 2015-19:

$D_{KL}(p_{ij}^{15-19} \| p_{ij}^{15-19}) = 0$. Note that the entropy of p_{ij}^{15-19} is not zero, but the entropy in the presence of prior knowledge is zero, i.e. the uncertainty remaining after the prior knowledge is accounted for is zero.

Although the dependence structure in international migration flows is relatively stable, more recent flows are better predictors of current flows than more distant flows. Table 3.8 shows the uncertainty remaining after knowledge of the migration flow in a given period. The uncertainty remaining is quantified by the KL information divergence. A high information divergence means a high remaining uncertainty.

Table 3.8 KL information divergence between historical migration flow and migration in 2015-19

Reference period	KL information divergence
1990-95	0.01252
1995-00	0.00833
2000-05	0.00484
2005-10	0.00234
2010-15	0.00065
2015-19	0.00000

4 Simulation of microsystems: random walk

4.1 Introduction

In the previous section, the focus was on the macrostates produced by individual preferences and immigration quota. Macrostates are described by population and transition counts. In this section, the attention shifts to microstates. In microstates individuals are uniquely identified by an identification number (ID). Individual differences are caused by observed individual heterogeneity and individual stochasticity due to intrinsic randomness (for the concept of individual variability, see Caswell, 2009). The study of microstates requires microdata, either observed data from censuses, population registers or sample surveys, or simulated data. In this section, the microdata are generated by a model the parameters of which are estimated from empirical data. Making inferences about individuals from aggregate data may lead to ecological fallacy. Relationships at the group-level do not automatically characterize the relationship at the level of the individual. Being mindful of population heterogeneity helps avoid the ecological fallacy (Courgeau et al., 2017; Bijak et al., 2018, p. 171). In this chapter, aggregate characteristics are not regarded as substitutes for individual characteristics. Instead, actors at a higher level of aggregation, in this case governments, impose conditions that define properties at the macrolevel and constrain actions of individuals.

Individual data with global coverage are rare. Sample surveys are generally organized at the country level. The World Fertility Survey and its offspring, the series of Demographic and Health Surveys, the series of national Health and Retirement Studies (https://g2aging.org), and the International Household Survey Network (https://www.ihsn.org) cover large parts of the world. IPUMS (https://www.ipums.org) is a data archive of census and survey data from around the world. Combined the sample surveys approximate a global coverage. Migration surveys with a global coverage do not exist. They are in the stage of being proposed (Willekens et al., 2016; Cerrutti et al., 2021). The advantages of sample surveys for the study of international migration are known for decades (see e.g. Fawcett and Arnold, 1987). The World Gallup Poll, which includes

migration-related data, aims at a global coverage. In the absence of micro data, data on individuals are inferred from population-level data. It is a common practice to produce virtual populations. If data are combined from different sources, the virtual population is also known as synthetic population. The practice of constructing a virtual population from aggregate data is followed in this paper. A related practice is the sampling from contingency tables (Kyabi et al., 2018; DeSalvo and Zhao, 2020).

A virtual population is a computer-generated population. The population is not directly observed, but is generated by stochastic simulation using a probability model. The parameters of the model are estimated from empirical data, usually incomplete data. The aim is to produce a virtual population that statistically mimics, i.e. is as close as possible to, a real population. An advantage of the approach is that data from different sources may be combined to generate a virtual population, and that the virtual population may be used to study 'what if' questions. The approach may be viewed as an extension of imputation. The imputation of missing data infers information that is not observed, but predicted by a model. The result is an augmented reality, i.e. a reality augmented by model outcomes. The validity of the predictions is very much dependent on the assumed distribution of individual characteristics in a population. The inference is valid if the virtual population has the same characteristics and behaves similarly as the real population under investigation. Validity implies agreement between the simulated system (microstates) and the observed system (macrostates). The production of a virtual population has also much in common with modelling unobserved population heterogeneity, in which latent subject-specific effects are added to population-level effects. The distribution of subject-specific effects in a population is represented by probability distributions with parameters estimated from empirical data. The Gumbel function, covered in Chapter 2, is such a distribution. It describes the effects of unobserved attributes of individuals and places on the utilities individuals attach to places.

A particularly flexible individual-level model is the random walk. Individual moves are random, but they are constrained in different ways. In this paper, individual moves are constrained by governments acting on information on the collective behaviour of individuals. Constrained random walks have a long history. Metropolis et al. (1953) developed a method that constraints random walks and produces a collective outcome with desired properties. The method became known as the Random Walk Metropolis (RWM). It is one of the most common Markov Chain Monte Carlo (MCMC) algorithms in use today (see e.g. Sherlock et al., 2010). The method is very flexible. It enables extensions that consider personal attributes, individual differences in decision making and interactions between individuals. Casati et al. (2015) use the method to generate virtual (synthetic) populations in the presence of control totals (see also Ye et al., 2017; Müller, 2017). Yaméogo et al. (2021) present a state-of-the-art review of methods for constructing synthetic populations.

In this section, we use a sample of the world population. The sample size is one million individuals. Individual attributes and actions are inferred from probability distributions with parameters estimated from count data. Individual values are obtained by sampling from the probability distribution (probability mass functions since the personal attributes considered are discrete).

This section consists of four subsections. The second subsection is a brief note on sampling discrete distributions. The third briefly reviews different types of random walk and pictures migration as a

random walk with preferences and barriers. The final subsection is an application to the world system of six regions.

4.2 A note on sampling

Stochastic simulation, or Monte Carlo simulation, involves sampling from multivariate probability distributions. It involves drawing (pseudo-)random numbers from probability distributions. Each draw relates to a single individual. The outcome of the sampling is a virtual population of individuals with personal attributes determined by the random draws. The results of sampling are collected in tables of counts, i.e. contingency tables. In this paper, the probability distributions are discrete and joint distributions are constrained by given marginal distributions. Sampling from multivariate distributions with fixed marginal distributions is equivalent to sampling from contingency tables with given marginal totals (Dobra and Mohammadi, 2018). In case of discrete multivariate data (cross-classified count data), Bishop et al. (1975, pp. 62ff; pp.435ff) distinguish the following sampling distributions:

- *Independent Poisson sampling*: each cell has an independent Poisson distribution. By implication, the total sample size is not fixed.
- *Simple multinomial sampling*: when the total sample size is fixed, the set of independent Poisson distributions gives a multinomial distribution. A parameter of the multinomial distribution is the probability that a randomly selected individual falls in a given category, e.g. the probability that a randomly selected individual resides in region i.
- *Product multinomial sampling*: stratified simple random sampling with the total sample size in each stratum fixed. The sampling scheme is the product of simple multinomial sampling within strata. For instance, individuals are sampled from a population stratified by place of origin and place of destination.

Simple and product multinomial sampling are the dominant types of sampling used in this paper. An advantage of product multinomial sampling is that aggregation of simulated individual data reproduces the given contingency tables exactly. Simple multinomial sampling reproduces the contingency tables approximately. Poisson sampling is not used in this paper because we have no data on migrations (events) but on migrants (transitions) only. Poisson sampling produces data with a variance equal to the mean. In migration data, the variability is often larger than expected from the Poisson distribution, a phenomenon known as *overdispersion*. To account for the extra variability, the parameter of the Poisson distribution is considered to vary randomly, i.e. it is a random variable. If the parameter follows a gamma distribution, then the gamma-Poisson mixture distribution is a negative binomial distribution, with two parameters, the mean of the Poisson parameter and a shape parameter of the gamma distribution, which controls the deviation from the Poisson distribution[10]. Poisson sampling is very relevant in migration research, but data issues

[10] Suppose a migrant makes several attempts to enter a country before a successful entry. Assume each attempt is an independent Bernoulli trial with probability of success p and probability of failure 1-p. The number of attempts (failures) until success is a random variable with a negative binomial distribution. If the number of attempts without restriction follows a Poisson distribution with parameter λ, then λ follows a gamma distribution with shape parameter the number of successes r (r=1 in this example) and scale parameter $\theta = \frac{p}{1-p}$ with $p = \frac{\lambda}{r+\lambda}$. Hence $\lambda = \frac{pr}{1-p}$ and the variance is $\lambda\left(1+\frac{\lambda}{r}\right) > \lambda$. The smaller r, the larger the overdispersion. If $r \to \infty$, the negative binomial distribution converges to a Poisson distribution with variance equal to the mean. To sample a negative binomial distribution, simulate a sequence of

prevent its use in this paper. The available data are transition data, i.e. data on places of residence at two points in time. The total number of individuals is fixed and, consequently, the maximum number of relocations is fixed too. It calls for multinomial sampling.

4.3Migration as a random walk with bias and constraints

The system of regions considered in the previous sections consists of r region. The random variables ${}_kX(t)$ and ${}_kX(t+1)$ denote the region of residence of individual k at time t and t+1, respectively. The previous sections emphasized the probability distributions of possible values of ${}_kX(t)$ and ${}_kX(t+1)$ (state probabilities) and the joint distributions of ${}_kX(t)$ and ${}_kX(t+1)$ (transition probabilities). The emphasis on probability distributions of random variables rather than on individual values of the random variables lead to a description of macrostates. In this section, the emphasis is on individual values of random variables and microstates.

The sequence of regions of residence of individual k is $\{ {}_kX(t),\ {}_kX(t+1),\ {}_kX(t+2), \ldots \}$. The sequence is a stochastic process, generally known as a *random walk*. A stochastic process has many possible realizations. A particular realization is the *sample path* of the stochastic process.

Dividing the world in a system of regions makes space discrete and the random walk a sequence of regions. A relocation is a step in a random walk. A random walk may also permit the individual to stay in the current location. Such a random walk is known as a *lazy random walk*. The destination of a step is random and the possible destinations follow a probability distribution. In a simple random walk, all possible destinations are equally probable. If some destinations are more probable than other, the random walk is referred as a *biased random walk*. In this paper, the random walks are mostly lazy and biased.

The sequence of locations may be recorded without or with the time of relocation. In a recent review, Dshalalow and White (2021) refer to time-insensitive and time-sensitive random walks. If time is disregarded, the position of a walker is recorded after each step. If the walker lacks memory, a destination is independent of previous locations and the random walk is a Markov chain. The sequence of steps is governed by relocation probabilities. If the time of relocation matters and the walker lacks memory, then the random walk is a Markov process. Time can be discrete (divided into time intervals) or continuous. Random walks in discrete time are governed by relocation probabilities. If time is continuous, the walk is known as a *continuous-time random walk* (CTRW) and is governed by relocation rates. In the absence of memory, the CTRW is a continuous-time Markov process. In a CTRW, both time and destination are random variables. The first is described by a waiting-time distribution and the second by a Markov chain. The CTRW may be extended further by distinguishing process time (duration in current location) and physical time (time since the start), which introduces age in random walks, as recently shown by Giona et al. (2019). There are indications that the formalism of a random walk is tending towards that of continuous-time Markov processes and semi-Markov processes. These stochastic processes are also the probabilistic foundations of multiregional demography. Random walks in a system of regions therefore represent the extension of multiregional models to individual-based models (IBM). Multistate microsimulation models in continuous time may be viewed as a continuous-time

---

Bernoulli trials with parameter of success p by sampling values u from a standard uniform distribution U(0,1). When the number of cases with $u < p$ reaches r, the number greater than p (failures) is a negative binomial random number.

(https://en.wikipedia.org/wiki/Negative_binomial_distribution)

random walk with transition intensities dependent on personal attributes and other predictors used in the model. Illustrations include microsimulation model published by Zinn (2014), the agent-based model by Klabunde et al. (2017) and its implementation in the domain-specific language ML3 (Warnke et al., 2017) and its extension by Reinhardt et al. (2022).

Approaching a random walk as a Markov chain and a Markov process has several advantages, one being that models of multiregional demography can be applied. Multiregional demography studies the dynamics of a population in a system of regions. Random walks offers the opportunity to introduce individual behaviour into multiregional population dynamics. Approaching migration as a random walk is not new. Yasuda (1975) describes migration as a random walk. Zwanzig (1983) notes that migration between regions can be treated as a CTRW between these regions. The author also used the equivalence between the CTRW and the master equation, which is a flow equation used in physics and which resembles the flow equation used in multiregional demography. Weidlich and Haag (1988) made extensive use of the master equation in the study of migration, but they did not make the connection with CTRW. Kanaroglou et al. (1968) and Haag (2017) use random utility theory to extend the master equation to a choice process. The CTRW is an active field of research. Giona et al. (2019) and Dshalalow and White (2021) review the state of the art. Allegrini et al. (2003) added age to the Zwanzig model, turning the CTRW into a demographic model. Relocation in the Schelling model has also been described as a random walk (Shin and Sayama, 2014). Barbosa et al. (2018) and Riascos and Mateos (2021) illustrate the growing popularity of the random walk to model mobility patterns. Riascos and Mateos consider a biased random walk with preferential navigation. Preferences are determined by features of nodes in the network, very similar to place utility concept used in this paper. The proliferation of tracking technologies and digital mobility data contribute to that development in the modelling of mobility (Luca et al., 2023). Of importance is also the gradual shift from spatial configurations (system of regions/places) to a network configuration (network of regions/places). It paves the way to network analysis of spatial mobility patterns (see also Abel et al., 2021; Bijak, 2021). Abel et al. treat the world as a single network and propose entropy maximization and a random walk algorithm to find clusters of interconnected countries in the network.

A random walk may be restricted by the presence of barriers. When the walker reaches a barrier, the walk may end (*absorbing barrier*) or may induce a return to the previous position (*reflecting barrier*). Other responses are possible too. A particularly important constraint, considered in this paper, consists of a threshold or capacity constraint. It is the maximum number of individuals in a place, e.g. country, at a given point in time. Immigration quota is a capacity constraint. A variety of restrictions may be imposed on the individual random walk. Steps in the random walk are conditional on meeting the constraints. If a step implies that a constraint is violated, the step is not permitted. Instead of constraints, a collective goal may be imposed. For instance, an individual step is allowed if it increases the overall utility or level of satisfaction in a population. In estimation problems, an individual move is admitted if it makes the collective behaviour (macrosystem) more probable, given the information available on the collective bahaviour. Moves that do not contribute to a more probable macrostate are not admitted or are admitted with a certain probability only. Hence individual moves that contribute to a more probable macrosystem are sampled more frequently than moves that do not contribute to the goal. This type of sampling is *acceptance-rejection sampling*. It forms the basis for algorithms such as the Metropolis algorithm.

Destination preference results in biased random walks. Recall that the odds that k prefers j rather than the reference region r' is given in equation (2.17):

$$ {}_k\theta_j = \frac{{}_kp_j}{{}_kp_{r'}} = exp\left[\ {}_kv_j - \ {}_kv_{r'}\right] $$

The multinomial logit is given in (2.18). The probability that k selects j as the preferred region of residence is given by (2.16). In the absence of data on place utilities, the location preferences are the preferences revealed by the migration flows during a past reference period. In Section 2, the revealed preference is by $p_{ij}^0 = n_{ij}^0 / \sum_j \ n_{ij}^0$ with n_{ij}^0 the observed number of individuals with residence in i at the beginning of the reference period and residence in j at the end of the reference period. The probability that a residents of region i prefers j over other regions is equal to the proportion of residents of i at the beginning of the reference period that reside in j at the end of the reference period. The equation is valid only if (a) individuals are allowed to stay in their region of residence and (b) the same distribution of destination preferences applies to all residents of i. Hence the probability that individual k selects j as the preferred region of residence is:

$$ {}_kp_{ij}^0 = p_{ij}^0 = n_{ij}^0 \ / \ n_{i+}^0 \qquad for\ all\ k \tag{4.1} $$

To determine which destination individual k with current residence i chooses as the preferred region of residence, two procedures may be followed. The first involves a Bernoulli trial, the second a random draw from a multinomial distribution. Let ${}_ku$ denote a (pseudo)random number drawn from a standard uniform probability distribution. If ${}_ku \leq p_{ij}^0$, then individual k prefers region j, otherwise k does not prefer region j. The outcome of the simulation experiment is a Bernoulli random variable, which is 1 of k prefers j over other regions and 0 otherwise. The experiment is equivalent to a single random draw from a Bernoulli distribution. The second procedure is to drawn a random number from the multinomial distribution with parameters $\{p_{i1}^0, p_{i2}^0, p_{i3}^0, \dots, p_{ir}^0\}$ [11]. The sampling results in a random vector of length r with elements 0 except for the element that denotes the preferred destination, which receives a value 1. Notice that the procedure implies that individuals in region i differ randomly in the place utility they attach to region j and that the differences follow a Gumbel distribution (See Section 2).

If all individuals in i have the same preferences $\{p_{i1}^0, p_{i2}^0, p_{i3}^0, \dots, p_{ir}^0\}$, a shortcut is to sample n_{i+} values from the multinomial distribution, with n_{i+} the number of residents in i at a point in time. Sampling generates a vector that gives the number of residents of i by preferred region of residence (including the current location). It does not identify the individuals who prefer a given region, j say. The reason in the absence of IDs. To determine the IDs of the individuals who prefer j, $p_{ij}^0 n_{i+}$ individuals are selected at random from the n_{i+} residents of i, without replacement. The procedure is repeated for all j. Suppose sampling starts with sampling $p_{i1}^0 n_{i+}$ individuals from all residents of i. At the start of the procedure, no resident of i has been assigned a location preference. After the sampling, $p_{i1}^0 n_{i+}$ residents of i received a location preference. The selected individuals prefer region 1. To determine the individuals who prefer region 2, $p_{i2}^0 n_{i+}$ individuals are selected at random from the residents of i who did not yet receive a location preference. The procedure is repeated until all n_{i+} individuals have a location preference assigned. Note that, since $\sum_{j=1}^{r} p_{ij}^0 = 1$, exactly n_{i+} individuals are assigned a location preference. The computational procedure is flexible. It allows for more complicated sampling than simple random sampling. For instance, individuals born in region h may be assigned much higher preferences for that region than assigned to individuals not born in h.

[11] An alternative but equivalent method is a random draw from the standard uniform distribution. If the value is between 0 and p_1, the destination is 1. It is 2 if the value is between p_1 and p_1+p_2. It is 3 if the value is between p_1+p_2 and $p_1+p_2+p_3$, and so on.

To accommodate immigration quota and other restrictions, a step in the random walk involves two activities. A first activity is to propose a preferred destination. If the proposal is accepted, it is followed by the actual relocation. Two variants of a random walk are considered: (a) biased and restricted random walk, and (b) biased and restricted random walk with an additional acceptance criterion. The first variant is covered in this subsection, the second is not covered in this paper.

For each member of the population a single random number is drawn from a multinomial distribution with parameters $\{p_{i1}^0, p_{i2}^0, p_{i3}^0, \dots, p_{ir}^0\}$, where i is the current location of an individual and $\sum_{j=1}^{r} p_{ij}^0 = 1$. Most residents of i prefer to stay in i. Suppose resident k proposes a relocation to region j. The proposal is accepted if the total number of individuals who already moved to j, which we denote by n_{+j}^*, is strictly less than the immigration quota imposed, i.e. $n_{+j}^* < n_{+j}$. The condition is a capacity constraint. Regions that have not reached their immigration quota are said to have vacancies, a concept borrowed from the Schelling model. The acceptance criterion is therefore $n_{+j}^* + 1 \leq n_{+j}$. If region j has no vacancies, then k may update her location preferences and propose an alternative destination. Such a response introduces a substitution effect of the immigration quota imposed by j. The multinomial distribution is adjusted by excluding j from the set of possible destinations (choice set). Let R^* denote the subset of destinations h for which $n_{+h}^* \leq h$. To determine the destination of an individual currently in i a random number is drawn from a multinomial distribution with parameters $p_{ih}^* = \frac{p_{ih}}{\sum_{m \in R^*} p_{im}}$ and $h \in R^*$. An alternative approach is to replace p_{ij}^0 by zero with j a region that reached its capacity constraint, and update the preferences such that $\sum_{i=1}^{r} p_{ij}^0 = 1$. A number is drawn from the updated multinomial distribution and the procedure is repeated until each individual in the population is able to move to a region with vacancies. Note that the move is not to the initial preferred region of residence, but to another region that became a preferred region after the regions that reached their immigration quota are removed from the set of alternatives.

Two remarks are in order. The first relates to the fairness of the procedure and the second to the extent to which the procedure reflects individual destination preferences.

- The migration flows generated by this procedure depend on the order in which individuals are selected. If all individuals in a given origin are selected first, they do not face capacity constraints and their relocations at the end of procedure are fully determined by their location preferences. Individuals selected later in the procedure are much more likely not to be able to move to their preferred location because of capacity problems. The procedure is unfair because the sequence in which individuals are selected determines the probability of facing capacity constraints. To make the procedure fair, individuals are ordered randomly before the start of the procedure. In other words, a queue of individuals is constructed with positions in the queue determined at random. Random ordering makes that each individual is equally likely to be early in the queue. Every member of the population has the same probability to move to a destination in accordance with the given destination probabilities.
- The procedure results in a migration flow matrix that satisfies the given marginals, but the flow is not necessarily the most probable flow given the marginal totals and the location preference matrix. The interaction between current region of residence (origin) and new region of residence (destination) is not necessarily identical to the dependence structure in the matrix of location preferences due to randomness. In other words, the individual actions may generate a flow matrix (macrosystem) that deviates from the one produced by

information gain minimization (KL method) and the iterative proportional fitting procedure. That is precisely what is observed in the application presented in subsection 4.4.

4.4 Application

The application consists of two steps. A first step is to construct a virtual population by assigning attributes to individuals. The second step is to simulate the migration behaviour of individuals in the presence of immigration restrictions. The construction of a virtual population is described in Annex C. The result is a person data structure (data frame or person data file). This section concentrates on the simulation. The individual-level results of the simulation are added to the data frame. After completing the simulation, the individual data can be analyzed using methods developed for the analysis of sample surveys, with the remark that the variance is caused by the simulation (Monte Carlo variance) and not by observed differences between individuals. The individual data represent the microsystem in which each individual is uniquely identified. The macrosystem is described by count data presented in tabular form or contingency tables.

In the absence of immigration restrictions (quota), individuals migrate to their preferred regions of residence. As a result, everyone is satisfied at the end of the interval (20). The number of residents of region i at t (2015) that resides in region j at t+1 (2020) is fully determined by the individual preferences and randomness. The individual location preferences are assumed to be consistent with the preferences revealed in the 1995-2000 migration flow. Table 4.1 shows the number of individuals in the virtual population by region of residence at t (2015) and region of residence at t+1 (2020) in the absence of immigration restrictions. The flow matrix is equal to the matrix obtained by multiplying the sample population by region of residence in 2015 and the preference matrix, except for small differences due to randomness associated with random sampling from the available contingency tables. The figures in Table 4.1 are comparable to the figures in Table 3.1 in Section 3, but two differences exist. First, the figures in Table 3.1 are population figures. They refer to the world population in 2015. The figures in Table 4.1 relate to a sample of the world population. Second, the figures in Table 3.1 are expected values, whereas the figures in Table 4.1 are sample counts. They converge to the expected values when the sample size becomes very large.

Table 4.1 Sample population in 2020, by region of residence in 2015, based on location preferences

	Region of residence in 2020						
Residence2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66330	267	77	94	191	167	67126
USCan	118	46094	268	14	176	18	46688
LatAm	116	1005	83300	1	26	2	84450
Africa	456	154	2	172064	136	24	172836
Asia	411	663	27	55	593488	571	595215
Rest	272	54	4	3	224	33128	33685
Sum	67703	48237	83678	172231	594241	33910	1000000

Let's introduce immigration quota. The quota is the difference between the sample population in 2020 (the column totals in Table C.3) and the number of stayers based on location preferences (diagonal elements of Table 4.1). Table 4.2 summarizes the results obtained until now. The first columns show the sample population in 2015, present in 2020, and the sample population in 2020. The number satisfied in 2015, by region of residence, is shown in column 3. The next two columns show the number dissatisfied, first by region of residence in 2015 (dissat015) and next by region of

preference (dissatD15). Note that globally the proportion of the population that is dissatisfied with their current region of residence on 0.560 percent (5596/1000000), which is very similar to the expected value shown in table 3.2 (0.570 percent). The immigration quota are shown next. The number of individuals selected for immigration and admitted in their preferred region (nselectedD), is the either the number dissatisfied (dissatD15) or the immigration quota, whatever is lowest. The next columns show the number of individuals selected and not selected, by preferred region of residence. The last column indicates the size of the unfilled quota. Globally, 83 percent of the applications for admission are accepted. Three regions do not reach their immigration quota as computed in this section: EU+, Latin America and the Caribbean, and Africa.

Table 4.2 Summary of results: sample population

	pop2015	pop2020	satisfied15	satsified19	dissatO15	dissatD15	quota	selectedD	unfilledquota
EU+	67127	67898	66288	66330	796	1373	1568	1373	195
USCan	46688	47458	46083	46094	594	2143	1364	1364	0
LatAm	84451	84115	83318	83300	1150	378	815	378	437
Africa	172836	172538	172040	172064	772	167	474	167	307
Asia	595212	594095	593442	593488	1727	753	607	607	0
Rest	33686	33896	33131	33128	557	782	768	768	0
Sum	1000000	1000000	994302	994404	5596	5596	5596	4657	939

The admission probabilities are shown in Table 4.3. They are based on the number of *dissatisfied* people in the sample population. They differ slightly from those in Table 3.3 due to sample variation. Table 4.4 shows the number of individuals whose application for admission are accepted/rejected, by region of origin and preferred region of residence, sample population.

Table 4.3 Admission probabilities in sample population

	Preference					
IDc	EU+	USCan	LatAm	Africa	Asia	Rest
EU+	0.00000	0.12459	0.20370	0.56287	0.25365	0.21355
USCan	0.08594	0.00000	0.70899	0.08383	0.23373	0.02302
LatAm	0.08449	0.46897	0.00000	0.00599	0.03453	0.00256
Africa	0.33212	0.07186	0.00529	0.00000	0.18061	0.03069
Asia	0.29934	0.30938	0.07143	0.32934	0.00000	0.73018
Rest	0.19811	0.02520	0.01058	0.01796	0.29748	0.00000
Sum	1.00000	1.00000	1.00000	1.00000	1.00000	1.00000

Individuals who are admitted in their preferred region of residence migrate to that region, while individuals whose application for admission is rejected stay in their initial region of residence (and remain dissatisfied with their place of residence). The distribution of the population by initial region of residence (2015), region of residence at the end of the procedure, and level of satisfaction at that time is shown in Table 4.5. Notice that 34 individuals in USACan are dissatisfied. The reason is that they were dissatisfied with USACan in 2015 and preferred to relocate to Asia, but their application was rejected (see Table 4.4, panel B).

Table 4.6 shows the population by region in 2015 and 2020. The difference between the 'observed' sample population in 2020 in EU+ (67898) and the population predicted by the model (67840) is 58. It is the difference between the unfilled quota (195; Table 4.2) and the number of applications by EU+ residents in 2015 rejected (137; Table 4.5).

Table 4.4 Number of individuals whose application for admission are accepted/rejected, by region of origin and preferred region of residence, sample population.

D. Application for admission granted

CurrentRegion \ PreferredRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0	170	77	94	154	164	659
USCan	118	0	268	14	142	18	560
LatAm	116	640	0	1	21	2	780
Africa	456	98	2	0	110	24	690
Asia	411	422	27	55	0	561	1476
Rest	272	34	4	3	181	0	494
Sum	1373	1364	378	167	608	769	4659

E. Application for admission rejected

CurrentRegion \ PreferredRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0	97	0	0	37	3	137
USCan	0	0	0	0	34	0	34
LatAm	0	365	0	0	5	0	370
Africa	0	56	0	0	26	0	82
Asia	0	241	0	0	0	10	251
Rest	0	20	0	0	43	0	63
Sum	0	779	0	0	145	13	937

F. Total applications for admission

CurrentRegion \ PreferredRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0	267	77	94	191	167	796
USCan	118	0	268	14	176	18	594
LatAm	116	1005	0	1	26	2	1150
Africa	456	154	2	0	136	24	772
Asia	411	663	27	55	0	571	1727
Rest	272	54	4	3	224	0	557
Sum	1373	2143	378	167	753	782	5596

Table 4.5 Sample population by region of residence in 2015, region of residence in 2020, and level of satisfaction with their location in 2020

A. Satisfied

InitialRegion \ CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66330	170	77	94	154	164	66989
USCan	118	46094	268	14	142	18	46654
LatAm	116	640	83300	0	21	2	84079
Africa	456	98	2	172064	110	24	172754
Asia	411	422	27	55	593488	561	594964
Rest	272	34	4	3	181	33128	33622
Sum	67703	47458	83678	172230	594096	33897	999062

B. Not satisfied

InitialRegion \ CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	137	0	0	0	0	0	137
USCan	0	34	0	0	0	0	34
LatAm	0	0	371	0	0	0	371
Africa	0	0	0	82	0	0	82
Asia	0	0	0	0	251	0	251
Rest	0	0	0	0	0	63	63
Sum	137	34	371	82	251	63	938

C. Total

InitialRegion \ CurrentRegion	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66467	170	77	94	154	164	67126
USCan	118	46128	268	14	142	18	46688
LatAm	116	640	83671	0	21	2	84450
Africa	456	98	2	172146	110	24	172836
Asia	411	422	27	55	593739	561	595215
Rest	272	34	4	3	181	33191	33685
Sum	67840	47492	84049	172312	594347	33960	1000000

Table 4.6 Population by region in 2015 and 2020 (in million)

Residence2015 \ Residence in 2020	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66194	165	116	171	260	221	67127
USCan	178	45765	376	43	271	56	46688
LatAm	202	632	83589	2	17	10	84451
Africa	349	145	3	172196	126	17	172836
Asia	738	707	28	115	593178	447	595212
Rest	237	44	4	11	245	33145	33686
Sum	67898	47458	84115	172538	594095	33896	1000000

5 Conclusion and ways forward

The paper presents an actor-based multilevel multiregional model. Individuals and governments are the actors. They interact directly and indirectly. Individuals have agency; they can act on their preferences. Preferences are based on the subjective utilities individuals attach to places. The concept of place utility is central in this paper. Individual actions are constrained by governments acting on behalf of nation states. Nation states restrict where individuals can settle. Restrictions are imposed on collective phenomena, more specifically on number of persons admitted to settle in a country in a given time interval. The cap on immigration and eligibility criteria, if any, restrict the freedom of movement. In the paper, restrictions are not personalized and do not target people with certain attributes. Chance determines who is entitled to move to the preferred place of residence and who is not. Most immigration restrictions, e.g. visa policies and residence permits, differentiate between individuals on the basis of personal attributes.

Emigration decision making is a complex process in time and space. In this paper, the emphasis was on space. The time dimension was simplified significantly. Klabunde et al. (2017) and Willekens (2017) concentrate on the time dimension and disregard space (Willekens) or limit space (Klabunde et al.). Klabunde et al. simulate multi-stage emigration decision processeses embedded in the life course and in a historical context, i.e. in two time scales: age and calendar time. Stages of life and life events influence the stages of the decision process differently, and the influence varies in calendar time due to changes in the socio-economic and political context. The dependencies are estimated from migration survey data (Survey on Migration between Africa and Europe (MAFE), Beauchemin, 2018), augmented by census data and life history data from the Demographic and Health Survey. The migration decision model was implemented in NetLogo and the demographic events ware simulated in R, using the MicSim package (Zinn, 2014). Willekens (2017) approached the emigration decision process as a multistage stochastic process with competing risks, but did not embed the process in the life course. The age structure of emigration and ages-specific emigration rates are outcomes of the model. They are determined by ages at onset of the process and the rates of transition between stages. In both publications, the decision process being simulated is rooted in the theory of planned behaviour (Fishbein and Ajzen, 2010), in response to the frequent call in the agent-based modelling literature for more sound behavioural theories. The basic structure of the model is very similar to the 'horse race' random utility model (Marley and Colonius, 1992), which is an extension of the discrete choice random utility model and accounts for the time individuals take to accumulate and process evidence in favour of an alternative. The time, known as response time, deliberation time and decision time, is random and follows a waiting time distribution. A further extension of the model to incorporate space is likely to lead to the continuous-time random walk model (CTRW) discussed briefly in this paper.

Migration flows between places of origin and places of destination are outcomes of individual decision processes and policies enacted by nation states. The modelling of the decision process is inspired by the Schelling (1971, 2006) model. Individuals assign utilities to alternative places of residence depending on how well place characteristics reflect subjective desires and aspirations. Place utilities are continuous random variables. The probability distribution of possible values determines the individual location preferences and ultimately the preferred place of residence. Individuals who are dissatisfied with their place of residence desire to move. In the Schelling model and in the model presented in this paper, capacity constraints limit the freedom of movement. In the Schelling model, capacity is measures as number of vacancies. In this paper, capacity is expressed as immigration quota. The restrictions limit an individual's capability to achieve personal aspirations. The model may be approached as a particular operationalization of

the capability theory of Sen and Nussbaum incorporating the two basic clusters: individual quality of life and social justice. A system if considered fair if all eligible candidates have the same probability of being selected. Fairness is operationalized by a random selection of applicants. The model captures the basic features of the visa lottery systems that exist in some countries. Individuals who are not satisfied with their place of residence and are not able to act on their preferences may adapt their preferences and apply next time for admission in a country unfilled quota. This simple mechanism produces the substitution effects, which receive much attention in the literature and are viewed as unintended side effects of restrictive immigration policies. Individuals with applications for admission rejected are, however, also at risk of exploring alternative routes to make it to their preferred country of residence. Alternative routes, such as irregular migration, are beyond the scope of this paper, but the model's open design and its modular structure offer an opportunity for add-ons.

Migration is complex and uncertain. The model accounts for the uncertainties by approaching migration between places as an outcome of a stochastic process, more particularly a biased random walk with preferences and barriers. In essence, the random walk model is an individual-based spatial interaction model in which relocations are random, while the distribution of relocations and the position of actors at a point in time follow certain probability distributions. The model distinguishes population-level outcomes (macrosystem), which emphasize expected values, and individual-level outcomes (microsystem), which emphasize individual idiosyncrasy and variability (*sample path*). The challenge is to reduce the uncertainty in the estimates of migration by supplying relevant information. In the paper, two types of information on migration flows are distinguished. The first is information about the true flows. The information is usually not detailed but limited to aggregate data, e.g. total numbers of people by place of residence at two points in time. The second type is auxiliary information that is relatively detailed but is not directly about the true flow. It may consist of data on the distribution of location preferences or migration intentions in a population, on past migration flows, or on expert opinions or crowdsourced knowledge about migration flows. In the paper, individual location preferences revealed by past migration flows are used in the prediction process. When combining data from different sources, e.g. data on the true flows and auxiliary data, it is useful to measure the information content of each source of information and to determine the knowledge contributed by each piece of data. This is done by formulating the estimation problem as a mathematical programming problem, more particularly a constrained optimization problem. The knowledge about the true flow enters the optimization problem as constraints. The necessary knowledge not included in the constraints is retrieved from the auxiliary data. In this paper, the knowledge about the true flow is limited to the population distribution at time t and the immigration quota imposed during the period from t to t+1. Information on the dependence between origin and destination is derived from the revealed location preferences. It assumes that (a) individuals act on their preferences, but within the constraints imposed by governments and (b) location preferences are relatively stable. For testing the model and documenting the information transfer from input data to estimates, the number of stayers and the number of people in each region at t+1 are assumed to be known too.

The analysis reveals, as expected, that (a) the degree at which individual preferences are reflected in the estimated migration flow depends on the constraints and (b) the accuracy of the estimates depends on how well the individual location preferences capture current associations between origin and destination. Because of the constraints, absolute location preferences are not relevant. Even the relative location preferences are in some cases not relevant. What matters are the ratios of relative location preferences, i.e. the relative location preferences of residents of region i divided by the relative location preferences of residents of region j. The ratio of relative location preferences is

an odds ratio. It means that, in the presence of immigration restrictions, the capacity of individuals to act on their preferences depends on the location preferences of other individuals in the system of regions. The collective constraints imposed by nation states affect everyone's ability to act on individual preferences. An individual's ability to act on preferences depends on other people's actions. The individual freedom depends on the collective behaviour. The random mechanisms (lotteries) incorporated in the model ensure that everyone is affected equally by the restrictions imposed by governments or generated by the collective behaviour.

In the paper, much emphasis was placed on the spatial dimension and on the consistency between the actor-based model and the spatial interaction models that are currently used to estimate migration flows from incomplete data. The spatial interaction models are given a statistical foundation showing the equivalence between the modelling of spatial interactions and the modelling of contingency tables and the decomposition of the effects of origin and destination into main effects and interaction effects through log-linear modelling. The main effects can be attributed mainly to the knowledge of the population at t and the immigration quota during the period from t to t+1. The interaction effects are 'borrowed' from the auxiliary data, in this case the revealed location preferences. The strategy to derive an actor-based model from established spatial interaction models has been successful. The parameters of the model are estimated from aggregate data by sampling contingency tables, a technique that is often used to produce individual data in the absence of micro data.

For illustrative purposes, a system of six regions is considered and the migration flows between all countries of the world are aggregated into flows between six regions. The model may be applied to any system of regions, including a system consisting of all countries of the world. In such a system of regions, many flows are zero or practically zero. These zero values should be treated as sample zeros and not as structural zeros. The distinction between these two types of zeros in contingency tables is important because the estimation method preserves structural zeros, but does not (and should not) preserve sample zeros. The actor-based model may be extended in several ways:

a. Personal attributes may be added, e.g. skill level (Willekens, 2017) and migration may be embedded in the individual life course to account for the impact of life events, such as changes in marital status, employment status and health status on desires and intentions to migrate (Klabunde et al., 2017).
b. Actors may be given an opportunity to interact. They may form social, economic or political networks that influence future migration. Since in the paper country of birth and country of residence are included in the micro data on the virtual population, diaspora may be generated. If the life course perspective is adopted, kin and family networks (linked lives) may be constructed. They are particularly relevant in modelling transnational family formation and family reunion and the migrations they often trigger. Nation states may create political unions with freedom of movement internally and more restrictions on immigration from outside the union.
c. Sample migration survey data may replace the sampling of probability distributions and contingency tables, which is only an intermediate step of empirically sound actor-based modelling and simulation. Cross-sectional surveys exist for some parts of the world and the Gallup World Poll is a worldwide survey covering more than 150 countries. A comprehensive actor-based model of migration not only benefits from data collection, but may also guide data collection, such as a world migration survey.
d. Revealed location preferences may be replaced by stated preferences, intentions and aspirations. In the absence of reliable migration flow data, the use of intentions is particularly popular in forecasting (Tjaden et al., 2019). The predictive performance of

intentions is limited, however. Turning intentions into actions requires resources, financial and social, and the capability to remove obstacles. In the migration literature, 'intention' is often used as an umbrella concept without the specificity needed to test whether intentions are good predictors of actions (for a discussion, see Migali and Scipioni, 2019; Aslany et al., 2021; Willekens, 2021). For decades, psychologists tried to unravel the link between intentions and actions. It motivated Ajzen (1985) to extend the Fishbein's theory of reasoned action into the theory of planned behavior by distinguishing between individual beliefs about the capability to remove obstacles and the actual degree of control over behavioural outcomes. For a discussion on why intention is often a poor predictor of behavior, see Ajzen (2020, pp. 320ff). Tjaden et al. (2019) test the link between intentions and behaviour using six consecutive years in the Gallup World Poll between 2010 and 2015. They found "a strong association between emigration intensions and recorded bilateral flows" (p. 36 and 43). Migali and Scipioni (2019), analyzing the same data, are less positive about the predictive performance of intentions and suggest not to rely on intentions but use information instead on whether individuals are past the intention stage (and the planning stage) and are actually *preparing* for migration. They conclude that "However, and especially for policymakers, this article suggests that the preparation for migration is the aspect where most of the attention should be focused if the research intention is to capture future migrants." The recommendation is consistent with the process perspective on emigration decision making. Decision is a staging process that takes time. Each stage is affected by many factors, personal and contextual. Early stages of the decision process, e.g. the attitude and the intention stages, have limited predictive power because individuals are unable to predict all possible intervening factors. The predictive power increases in later stages of the decision process because the intervening factors are more predictable. Klabunde et al. (2017) simulate the transition from attitude to intention to planning to preparation accounting for some important intervening factors. Discussions on the predictive performance of intentions and discussions in economics on stated versus revealed preferences motivated, for this paper, the choice for revealed preferences.

e. In the model presented in the paper, time is omitted or is discrete. Most decision-making processes take time, in particular when they lead to life choices that affect the entire remaining life course. Decision making also consists of stages (Willekens, 2021). An extension of the model should reflect the process character of decision making, distinguish between stages, and should allow for individual differences in the pace of decision making. Klabunde et al. (2017) and Willekens (2017) use a process theory of planned behaviour and waiting time distributions to model durations of stages. The features can be integrated into a continuous-time random walk.
f. The model in this paper is essentially a logit model or multinomial logistic regression model (with one explanatory variable: region of current residence) for reasons described in the paper. It predicts the probability that an individual who resides in i at t resides in j at t+1. It does not predict migration counts. In order to predict migration counts during a given period, the logit model should be replaced by a counting process model in continuous time. Unlike in the logit model, a counting process model allows for multiple events during a time interval. It relates each occurrence to an appropriate *exposure time*. The simplest counting process is the Poisson process. The Poisson regression model is widely used in migration research (e.g. Raymer et al., 2013).
g. The multilevel model with individuals and governments as the actors, has, by design, an important policy component. As demonstrated in the paper, the model could relatively easily be extended to a migration policy model. Since individuals respond to policies,

frequently in unforeseen ways, the model could shed light on unintended consequences of policies and their impact on the effectiveness of policies. The extension of the model to a full policy model and more particularly a strategic foresight model for forward-looking governance requires long-term objectives and short-term decisions and actions that are consistent with long-term objectives (European Commission, 2020). Migration policy models with strategic foresight are not new. They were developed at a time of rapid urbanization and governments aimed at using internal migration as a policy instrument to achieve a more balanced population distribution across its territory (Willekens, 1979). The policy model developed at that time is essentially a forecasting model turned into an optimization model. It includes target variables, e.g. desired end states, and control variables representing policy actions. The model determines the future trajectory of actions needed to reach long-term objectives. What the model did not consider was how the actors involved would respond to policy measures. Most policies were only partly effective because many actors responded not as policy-makers expected them to respond. The belief in the makeable society and central planning was severely affected. Agent-based models, such as the Schelling model, demonstrated the limited impact of centralized coordination mechanisms.

h. If the number of spatial units is large, a network approach may be preferred over a multiregional approach because a network approach is able to reveal interesting structural characteristics of migration networks (Abel et al., 2021; Nagurney and Daniele, 2021). Since network analysis makes intensive use of matrix calculus, the matrix formulations of multiregional models (see Sections 2 and 3) remain valid in network analysis.

Migration is complex and uncertain because of the many actors involved and the decisions actors must make in the absence of complete information and while facing many uncertainties. Multi-actor models help unravel the complexities and information theory help master uncertainties.

References

Abel, G.J. (2013). Estimating global migration flow tables using place of birth data. *Demographic Research* 28(18): 504–546. Doi: http://dx.doi.org/10.4054/DemRes. 2013.28.18.

Abel, G. J. (2018). Estimates of global bilateral migration flows by gender between 1960 and 2015. *International Migration Review,* 52(3):809-852 https://doi.org/10.1111/imre.12327.

Abel, G.J. and N. Sander. (2014). Quantifying global international migration flows. *Science* 343(1520): 1520–1522. Doi: http://dx.doi.org/10.1126/science.1248676.

Abel, G.J. and J.J. Cohen (2019). Bilateral international migration flow estimates for 200 countries. *Scientific Data*, 6(82):1-13. Doi: 10.1038/s41597-019-0089-3

Abel, G. and J.J. Cohen (2021). Bilateral international migration flows for 200 countries (1990-1995 to 2015-2020). https://figshare.com/articles/dataset/Bilateral_international_migration_flow_estimates_for_200_countries_1990-1995_to_2010-2015_/7731233?backTo=/collections/Bilateral_international_migration_flow_estimates_for_200_countries/4470464

Abel, G., J. DeWaard, J. Trang Ha and Z.W. Almquist (2021). The form and evolution of international migration networks 1990-2015. *Population, Space and Place*, 27:e2432. Doi: 10.1002/psp.2432

Agresti, A. (2013). Categorical data analysis. 3rd edition. Wiley.

Ajzen, I. (1985). From intentions to actions: A theory of planned behavior. In: J. Kuhl and J. Beckmann eds., *Action control: From cognition to behavior*. Berlin, Heidelber, New York: Springer-Verlag. (pp. 11-39).

Ajzen, I. (2020). The theory of planned behaviour: frequently asked questions. *Human Behavior and Emerging Technologies*, 2:314-324. DOI: 10.1002/hbe2.195

Akaike, H. (1973). Information theory as an extension of the maximum likelihood principle. In: B. N. Petrov and F. Csàki eds. *Second International Symposium on Information Theory* Budapest: Akademiai Kiado. pp. 267-281. Reprinted (with introduction by J. deLeeuw) in S. Kotz and N.L. Johnson eds. (1992) Breakthroughs in statistics, vol. 1. London: Springer, pp. 599-624. DOI: 10.1007/978-1-4612-0919-5

Allegrini, P., G. Aquino, O. Grigolini, L. Palatella and A. Rosa (2003). Generalized master equation via aging continuous-time random walks. *Physical Reviews E*, 68(056123):1-11. Doi: 10.1103/PhysRevE.68.056123

Anas, A. (1983). Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research*, 17B(1):13-23. Doi: 10.1016/0191-2615(83)90023-1

Aslany, M., J. Carling, M.B. Mjelva and T. Sommerfelt (2021). *Systematic review of determinants of migration aspirations.* QuantMig Deliverable 2.2. Southampton: University of Southampton. Available at https://www.quantmig.eu/res/files/QuantMig%20D22%202021-01-29.pdf

Azose, J.J. and A.E. Raftery. (2015). Bayesian Probabilistic Projection of International Migration. *Demography* 52: 1627 – 1650. Doi: http://dx.doi.org/10.1007/s13524- 015-0415-0.

Azose, J.J. and A.E. Raftery (2019). Estimation of emigration, return migration, and transit migration between all pairs of countries. PNAS (Proceedings of the National Academy of Sciences), 116(1):116-122. Doi: 10.1073/pnas.1722334116

Azose, J.J., H. Sevcíková and A.E. Raftery (2016). Probabilistic population projection with migration uncertainty. PNAS (Proceedings of the National Academy of Sciences), 113(23):6460-6465. Doi: 10.1073/pnas.1606119113

Bandura, A. (2006). Towards a psychology of human agency. P*erspectives on Psychological Science,* 1(2):164-180. DOI : 10.1111/j.1745-6916.2006.00011.x

Barbosa, H., M. Barthelemy, G. Ghoshal and others (2018). Human mobility: Models and applications. *Physics Reports*, 134:1)74. Doi: 10.1016/j.physrep.2018.01.001

Barslund, M., M. Di Salvo and L. Ludolph (2019). Can regular replace irregular migration across the Mediterranean? Brussels: Center for European Policy Studies. Available at

https://www.ceps.eu/ceps-publications/can-regular-replace-irregular-migration-across-the-mediterranean/

Barthel, F. and E. Neumayer (2015). Spatial dependence in asylum migration. *Journal of Ethnic and Migration Studies*, 41(7):1131-1151. Doi: 10.1080/1369183X.2014.967756

Bawden, D. and L. Robinson (2015). "A few exciting words": Information and entropy revisited. *Journal of the Association for Information Science and Technology*, 66(10):1965-1987. Doi: 10.1002/asi.23459

Beauchemin, C. (2018). Migration between Africa and Europe. Cham: Springer. Doi: http:// dx.doi.org/10.1007/978-3-319-69569-3.

Beaumont, M.A. (2019). Approximate Bayesian computation. *Annual Review of Statistics and its Application*. 6:379-4043. Doi: 10.1146/annurev-statistics-030718- 105212

Beine, M., S. Bertoli and J. Fernández-Huertas Moraga (2016). A practitioners' guide to gravity models of international migration. *The World Economy*, 39(4):496-512. Doi: 10.1111/twec.12265

Beine, M., M. Bierlaire and F. Docquier (2021). Nex York, Abu Dhabi, London or stay at home? Using a cross-nested logit model to identify complex substitution patterns in migration. Bonn: IZA – Institute of Labor Economics: Discussion Paper No. 14090. Available at: https://www.iza.org/publications/dp/14090/new-york-abu-dhabi-london-or-stay-at-home-using-a-cross-nested-logit-model-to-identify-complex-substitution-patterns-in-migration

Berlemann, M., E. Haustein and M.F. Steinhardt (2021). From stocks to flows – Evidence for the climate-migration-nexus. Bonn: IZA(Institute of Labor Economics). Discussion Paper No. 14450. Available at https://www.iza.org/publications/dp/14450/from-stocks-to-flows-evidence-for-the-climate-migration-nexus

Bertoli, S., J. Fenández-Huertas Moraga and L. Guichard (2020). Rational inattention and migration decisions. Journal of International Economics, 126(103364) :1-22. Doi : 10.1016/j.jinteco.2020.103364

Bijak J (2010) Forecasting International Migration in Europe: A Bayesian View. Dordrecht: Springer.

Bijak, J. (2021). Towards Bayesian mode-based demography. Agency, complexity and uncertainty in migration studies. Cham: Springer. Doi: 10.1007/978-3-030-83039-7

Bijak, J., D. Courgeau, R. Franck and E. Silverman (2018). Modelling in demography: from statistics to simulation. In: E. Silverman Methodological investigations in agent-based modelling. Cham: Springer, pp. 167-187. Doi: 10.1007/978-3-319-72408-9_9

Bijak, J., G. Disney, A.M. Findlay, J.J. Forster, P.W.F. Smith and A. Wisniowski (2019). Assessing time series models for forecasting international migration: Lessons from the United Kingdom. *Journal of Forecasting*, 38(5): 470-487. Doi: 10.1002/for.2576

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

Boltzmann, L. (1905). Über statistische Mechanik (on statistical mechanics). In: L. Boltzmann Populäre Schriften, Essay 19. pp. 308-363. Leipzig: Verlag von Johann Ambrosius Bart. Available at https://archive.org/details/populreschrifte00boltgoog (Internet Archive).

Carling, J. (2002). Migration in the age of involuntary immobility: Theoretical reflections and Cape Verdean experiences. *Journal of Ethnic and Migration Studies*,28(1), 5–42.

Casati, D., K. Müller, P.J. Fourie, A. Erath and K.W. Axhausen (2015). Synthetic population generation by combining a hierarchical, simulation-based approach with reweighting by generalized raking. *Transportation Research Record* 2493(1):107-116. Doi: 10.3141/2493-12

Caswell, H. (2009). Stage, age and individual stochasticity in demography. *Oikos*, 118:1763-1782. Doi: 10.1111/j.1600-0706.2009.17620.x,

Cerrutti, M., P. Fargues and M. Awambila (2021). The case for a world migration survey. Policy & Research Papers No. 25. Paris: International Union for the Scientific Study of Population. https://iussp.org/sites/default/files/PRP25.pdf

Chilton, R., and R. R. W. Poet. (1973). An entropy maximising approach to the recovery of detailed migration patterns from aggregate census data. *Environment and Planning A* 5:135–46. Doi: https://doi.org/10.1068/a050135

Cicalese, F., L. Gargano and U. Vaccaro (2019). Minimum-entropy couplings and their applications. IEEE Transactions on Information Theory. 65(6):3436-3451. Doi: 10.1109/TIT.2019.2894519

Clemens, M. and K. Gough (2018). Can regular migration channels reduce irregular migration? Lessons for Europe from the United States. CGD Brief Feb2018. Washington D.C.: Center for Global Development. Available at https://www.cgdev.org/publication/can-regular-migration-channels-reduce-irregular-migration-lessons-europe-united-states

Coffey, S., B.T. West, J. Wagner and M.R. Elliott (2020). What do you think? Using expert opinions to improve predictions of response propensity under a Bayesian framework. *Methoden-Daten-Analysen*, 14(2):1-40. Doi: 10.12758/mda.2020.05

Courgeau, D., J. Bijak, R. Franck and E. Silverman (2017). Model-based demography: Towards a research agenda. In: A. Grow and J. van Bavel eds. Agent-based modelling in population studies, pp. 29-51. Cham: Springer, pp. 29-51. DOI: 10.1007/978-3-319-32283-4_2

Cover,T.J. and J.A. Thomas (2006). Elements of information theory. Second edition. Hoboken, New Jersey: Wiley. ISBN-13: 978-0471241959

Czaika, M. and H. de Haas (2013). The effectiveness of immigration policies. *Population and Development Review*, 39(3):487-508. Doi: 10.1111/j.1728-4457.2013.00613.x

Czaika, M. and M. Hololth (2016). Do restrictive asylum and visa policies increase irregular migration into Europe? *European Union Politics*, 17(3):345-365. Doi: 10.1177/1465116516633299

Darroch, J.N. and D. Ratcliff (1972). Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5),1470-1480. DOI: 10.1214/aoms/1177692379

de Haas, H. (2021). A theory of migration: the aspirations-capabilities framework. *Comparative Migration Studies*, 9(8):1-35. Doi: 10.1186/s40878-020-00210-4

de Haas, H., M. Czaika, M.-L. Flahaux, E. Mahendra, K. Natter, S. Vezzoli and M. Villares-Valera (2019). International migration : trends, determinants, and policy effects. *Population and Development Review*, 45(4):885-922. Doi: 10.1111/padr.12291

Del Fava, E., A. Wiśniowski and E. Zagheni (2019). Modeling international migration flows by integrating multiple data sources. https://osf.io/preprints/socarxiv/cma5h/

Deming, W.E. and F.F. Stephan (1940). 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. DOI: 10.1214/aoms/1177731829

DeSalvo, S. and J. Zhao (2020). Random sampling of contingency tables via probabilistic divide-and-conquer. *Computational Statistics*, 35:837-869. Doi: 10.1007/s00180-019-00899-7

Dobra, A. and R. Mohammadi (2018). Loglinear model selection and human mobility. *The Annals of Applied Statistics*. 12(2):815-845. Doi: 10.1214/18-AOAS1164

Dshalalow, J.H. and R.T. White (2021). Current trends in random walks on random lattices. *Mathematics*, 9(1148):1-38. Doi: 10.3390/math9101148

Eshima, N. (2020). Statistical analysis and entropy. Singapore; Springer. Doi: 10.1007/978-981-15-2552-0

Etz, A. (2019). Technical notes on Kullback-Leibler divergence. https://psyarxiv.com/5vhzu/

European Commission (2020). 2020 Strategic Foresight Report. Strategic foresight – charting the course towards a more resilient Europe. COM(2020) 493 final. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020DC0493&from=EN

Fawcett, J.T. and F. Arnold (1987). The role of surveys in the study of international migration: an appraisal. *The International Migration Review*, 21(4):1523-1540. Doi:: 10.2307/2546523

Fienberg, SE. (1970). The iterative procedure for estimation in contingency tables. *The Annals of Mathematical Statistics*, 41(3):907-917. Doi: 10.1214/aoms/1177696968

Fienberg, S.E. (2006). When did Bayesian inference become "Bayesian"? *Bayesian Analysis*, 1(1):1-40. DOI: 10.1214/06-BA101

Fishbein, M. and I. Ajzen (2010). Predicting and changing behavior. The Reasoned Action Approach. New York: Psychology Press (Taylor and Francis). DOI: 10.4324/9780203838020

Flowerdew, R. and Aitkin, M. (1982). A Method of Fitting the Gravity Model Based on the Poisson Distribution. *Journal of Regional Science* 22: 191 – 202. Doi: http://dx.doi.org/ 10.1111/j.1467-9787.1982.tb00744.x.

Fotheringham, A.S. and M. Sachdeva (2021). Modelling spatial processes in quantitative human geography. *Annals of GIS*. Doi: 0.1080/19475683.2021.1903996

Geenens, G. (2020). Copula modelling for discrete random vectors. *Dependence Modelling*. 8(1):417-440. Doi: 10.1515/demo-2020-0022

Genest, C. (2021). A tribute to Abe Sklar. *Dependence Modeling*, 9:200-224. Doi: /10.1515/demo- 2021-0110

Gibbs, J.W. (1902). Elementary principles in statistical mechanics. New York: Schibner's Sons. eBook: Project Gutenberg (2016) available at http://www.gutenberg.org/files/50992/50992-pdf.pdf

Giona, M., M.D'Ovidio, D. Cocco,A. Cairoli and R. Klages (2019). Age representation of Lévy walks: partial density waves, relaxation and first passage statistics. *Journal of Physics A: Mathematics and Theoretical*. 52(384001):1-28. Doi: 10.1088/1751-8121/ab38eb

Good, I.J. (1950). Probability and the weighing of evidence. London: Griffin & Co.. ASIN: B0000CHL1R. Available at https://www.gwern.net/docs/statistics/bayes/1950-good-probabilityandtheweighingofevidence.pdf

Good, I.J. (1956). Some terminology and notation in information theory. *Proceedings of the IEEE – Part C. Monographs*, 103(3):200-204. DOI: 10.1049/pi-c.1956.0024

Good, I.J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables, *The Annals of Mathematical Statistics*,34(3):911-934. DOI: 10.1214/aoms/1177704014

Good, J.I. (1985). Weight of evidence: a brief survey. In: J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith eds. Bayesian statistics 2, Amsterdam: Elsevier, pp. 249-270. ISBN: 04448-87746-0

Haag, G. (2017). Modelling with the master equation. Cham: Springer. Doi: 10.1007/978-3-319-603001-1

Haghani, M., M.C.L. Bliemer and D.A. Hensher (2021). The landscape of discrete choice modelling research. *Journal of Choice Modelling*. 40(100303):1-40. Doi: 10.1016/j.jocm.2021.100303

Hegselmann, R. (2017). Thomas C. Schelling and James M. Sakoda: The intellectual, technical and social history of a model. Journal of Artificial Societies and Social Simulation. 20(3):15. DOI: 10.18564/jasss.3511

Hinsch, M. and J. Bijak (2022). Principles and state of the art of agent-based migration modelling. In. J. Bijak ed. Towards Bayesian model-based demography: Agency, complexity and uncertainty in migration studies. Springer. pp. 44-58.

Hitlin, S. and M.K. Johnson (2015). Reconceptualizing agency within the life course: The power of looking ahead. *American Journal of Sociology*, 120(5):1429-1472. DOI: 10.1086/681216

Holland, J.H. (1995). Hidden order. How adaptation builds complexity. Reading, Mass.: Helix Books.

Idel, M. (2016). A review of matrix scaling and Sinkhorn's normal form for matrices and positive maps. https://arxiv.org/abs/1609.06349

Ireland. C.T. and S. Kullback (1968). Contingency tables with given marginals. *Biometrika*, 55(1):179-188. DOI: 10.2307/2334462

Jaynes, E.T. (1957a). Information theory and statistical mechanics. *Physical Review*, 106(4): 620-630. DOI: 10.1103/PhysRev.106.620

Jaynes, E.T. (1957b). Information theory and statistical mechanics II., *Physical Review*, 108(2):171-190. DOI: 10.1103/PhysRev.108.171

Jaynes F.T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, 5 (33), pp. 391-398. DOI: 10.1119/1.1971557

Jaynes, E.T. (2003). Probability theory: the logic of science. (Annotated edition edited by G.L. Bretthorst). New York: Cambridge University Press. ISBN-13: 978-0521592710

Jeffreys, H. (1939). Theory of probability. Oxford: Oxford University Press (3rd edition 1961). ISBN-13 : 978-0198503682

Jeong, B., W. Lee, D.-S. Kim and H. Shin (2016). Copula-based approach to synthetic population generation. *PLoS One*. 11(8):e0159496. Doi: 10.1371/journal.pone.0159496

Jung, J., J.H. Kim, F. Matějka and A. Sims (2019). Discrete actions in information-constrained decision problems. *The Review of Economic Studies*, 86(6):2643-2667. Doi: 10.1093/restud/rdz011

Kanaroglou, P., K-L. Liaw and Y.Y. Papageorgiou (1986a). An analysis of migratory systems: 1. Theory. *Environment and Planning A*., 18:913-928.

Kanaroglou, P., K-L. Liaw and Y.Y. Papageorgiou (1986b). An analysis of migratory systems: 1. Operational framework. *Environment and Planning A*., 18:1093-1060.

Kass, R.E. and A.E. Raftery (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430):773-795. DOI: 10.1080/01621459.1995.10476572

Kateri, M. (2014). Contingency table analysis. Methods and implementations using R. New York: Springer. Doi: 10.1007/978-0-8176-4811-4

Kayibi, K.K., S. Pirzada and T.A. Chishti (2018). Sampling contingency tables. *AKCE International Journal of Graphs and Combinatorics*, 15:298-306. Doi: 10.1016/j.akcej.2017.10.001

Keynes, J.M. (1921). A treatise on probability. London: McMillan. Available at https://www.gutenberg.org/files/32625/32625-pdf.pdf

Klabunde, A. and F.J. Willekens (2016). Decision-making in Agent-based Models of Migration: State of the Art and Challenges. *European Journal of Population* 32(1): 73–97. Doi: https://doi.org/10.1007/s10680-015-9362-0.

Klabunde, A., S. Zinn, F. Willekens, and M. Leuchter (2017). Multistate Modelling Extended By Behavioural Rules: An Application to Migration. *Population Studies* 71(Supplement 1): S51–S67. Doi: https://doi.org/10.1080/00324728.2017.1350281.

Kullback, S.(1968). Probability densities with given marginals. *The Annals of Mathematical Statistics*, 39(4):1236-1243. DOI: 10.1214/aoms/1177698249

Kullback, S. and R.A. Leibler (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79-86. DOI: 10.1214/aoms/1177729694

Little, D.B. and R.J.A. Rubin (2020). Statistical analysis with missing data. 3rd edition. Hoboken, NJ: Wiley (first edition 1987).

Luca, M., G. Barlacchi, B. Lepri and L. Pappalardo (2023). A survey on deep learning for human mobility. *ACM Computing Surveys*, 55(1):article 7, pp. 1-44. Doi: 10.1145/3485125 (online November 2021)

Lusem, A.N. and M. Teboulle (1992). A primal-dual iterative algorithm for a maximum likelihood estimation problem. *Computational Statistics & Data Analysis*, 14:443-456. Doi: 10.1016/0167-9473(92)90060-S

Mackowiak, B., F. Matejka and M. Wiederholt (2022). Rational inattention: a review. *Journal of Economic Literature*, Forthcoming. Working Paper available at https://www.ecb.europa.eu/pub/pdf/scpwps/ecb.wp2570~a3979fbfa5.en.pdf

Mantzaris, A., J.A. Marich and T.W. Halfman (2018). Examining the Schelling model simulation through an estimation of its entropy. *Entropy*, 20(623):1-13. Doi: 10.3390/e20090623

Marley, A.A.J. and H. Colonius (1992). The "Horse Race" random utility model for choice probabilities and reaction times, and its competing risks interpretation. *Journal of Mathematical Psychology*, 36:1-20.

Massey, D.S., J. Arango, G. Hugo, A. Kouaouci, A. Pellegrino and J.E. Taylor (1993). Theories of international migration: A review and appraisal. *Population and Development Review*, 19(3):431-466. DOI: 10.2307/2938462

McAlpine, A., L. Kiss, C. Zimmerman and Z. Chalabi (2021). Agent-based modelling for migration and modern slavery research: a systematic review. *Journal of Computational Social Science*. 4:243-332. Doi: 10.1007/s42001-020-00076-7

McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In: P. Zarembka (ed.) Frontiers in Econometrics. New York: Academic Press, pp. 105–42. Available at https://eml.berkeley.edu/reprints/mcfadden/zarembka.pdf

McFadden, D. (1978). Modelling the choice of residential location. In: A. Karqvist, L. Lundqvist, F. Snickars and J. Weibull eds. Spatial interaction theory and planning models. Amsterdam: North Holland, pp. 75-96. Available at https://eml.berkeley.edu/reprints/mcfadden/location.pdf

Metropolis, N., A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller, and E. Teller (1953). Equations os state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1092. Doi: https://doi.org/10.1063/1.1699114.

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what to do about it. *European Sociological Review*, 26(1):67-82. Doi: 10.1093/esr/jcp006

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, 63(321):1-28. Doi: 10.2307/2283825

Müller, K. (2017). A generalized approach to population synthesis. PhD Thesis. ETH Zürich. Doi: 10.3929/ethz-b-000171586

Nagurney, A., P. Daniele and LS. Nagurney (2020). Refugee migration networks and regulations: a multiclass, multipath variational inequality framework. *Journal of Global Optimization*, 78:627-649. Doi: 10.1007/s10898-020-00936-6

Nagurney, A. and P. Daniele (2021). International human migration networks under regulations. *European Journal of Operations Research*, 291(3):894-905. Doi: 10.1016/j.ejor.2020.04.008

Napierala, J., J. Hilton, J.J. Forster, M. Carammia and J. Bijak (2021). Toward an early warning system for monitoring asylum-related migration flows in Europe. *International Migration Review*, Doi: 0.1177/01979183211035736

Nelson, R.B. (2006). An introduction to copulas. Second edition. New York : Springer. Doi : 10.1007/0-387-28678-0

Niven, R.K. (2009). Combinatorial entropies and statistics. *European Physical Journal B*, 70:49-63. Doi: 10.1140/epjb/e2009-00168-5

Norton,E.C. and B.E. Dowd (2018). Log odds and the interpretation of logit models. *Health Services Research*, 53(2):859-878. Doi: 10.1111/1475-6773.12712.

Nowok, B. (2010). Harmonization by simulation: a contribution to comparable international migration statistics in Europe. Amsterdam: Rozenberg Publishers. ISBN: 978-90-367-4549-9. Available at: https://research.rug.nl/en/publications/harmonization-by-simulation-a-contribution-to-comparable-internat

Nowok, B. and F. Willekens (2011). A probabilistic framework for harmonisation of migration statistics. *Population, Space and Place* 17: 521–533. Doi: https://doi.org/ 10.1002/psp.624.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician*. 73:supp1:69-81. Doi: 10.1080/00031305.2018.1518265

Okunade, S. (2021). Africa moves towards intracontinental free movement for its booming population. Migration Information Source. Washington D.C.: Migration Policy Institute. https://www.migrationpolicy.org/article/africa-intracontinental-free-movement

Ortega, F. and G. Peri (2013). The effect of income and immigration policies on international migration. *Migration Studies*, 1(1):47-74. Doi: 10.1093/migration/mns004

Pellegrini, P.A. and A.S. Fotheringham (2002). Modelling spatial choice: a review and synthesis in a migration context. *Progress in Human Geography*, 26(4):487-510. Doi: 10.1191/0309132502ph382ra

Plane, D.A. (1982). An information theoretic approach to the estimation of migration flows. *Journal of Regional Science*, 22(4):441-456. Doi: 10.1111/j.1467-9787.1982.tb00769.x

Potter, R.G. and J.M. Sakoda (1966). A computer model of family building based on expected values. *Demography*, 3(2):450-461. Doi: 10.2307/2060170

Pressé, S. (2013). Principles of maximum entropy and maximum caliber in statistical physics. *Review of Modern Physics*, 85(3):1115-1141. Doi: 10.1103/RevModPhys.85.1115

Raymer, J. (2007). The estimation of international migration flows: a general technique focused on the origin-destination association structure. *Environment and Planning A*. 39:985-995. Doi: 10.1068/a38264

Raymer, J., A. Wisniowski, J.J. Forster, P.W.F. Smith, and J. Bijak. (2013). Integrated Modeling of European Migration. *Journal of the American Statistical Association* 108(503): 801–819. Doi: https://doi.org/10.1080/01621459.2013.789435.

Raymer, J., F. Willekens and A. Rogers (2018). Spatial demography: a unifying core and agenda for further research. *Population, Space and Place*, 25(4):1-13 DOI: 10.1002/psp.2179

Raymer, J., Q. Guan and J. Trang Ha (2019). Overcoming data limitations to obtain migration flows for ASEAN countries. *Asian and Pacific Migration Journal*, 28(4):385-414. Doi: 10.1177/0117196819892344

Rees, Ph. and F. Willekens (1986) Data and Accounts. In: A. Rogers and F. Willekens eds. Migration and settlement: a multiregional comparative study. Dordrecht, The Netherlands: Reidel Press, pp. 19-58. Available at https://pure.knaw.nl/ws/portalfiles/portal/640378/18923.pdf

Reinhardt, O., T. Warnke and A.M. Uhrmacher (2022). Agent-based modelling and simulation with domain-specific languages. In. J. Bijak ed. Towards Bayesian model-based demography: Agency, complexity and uncertainty in migration studies. Springer. pp. 118-136.

Riascos, A.P. and J.L. Mateos (2021). Random walks on weighted networks: Exploring local and non-local navigation strategies. *Journal of Complex Networks*. 9(5):cnab032. Doi: 10.1093/comnet/cnab032

Rikani, A. and J. Schwede (2021). Global bilateral migration projections accounting for diasporas, transit and return flows, and poverty constraints. *Demographic Research*, 45(4):87-140. Doi: 10.4054/DemRes.2021.45.4

Robert, C.P. (2011). Reading Keynes' Treatise on Probability. *International Statistical Review*, 79:1-15. Doi: 10.1111/j.1751-5823.2010.00129.x

Rogers, A. (1995). Multiregional demography. Principles, methods and extensions. Chichester: Wiley.

Rogers, A., F. Willekens and J. Raymer (2003). Imposing age and spatial structure on inadequate migration flow datasets. *The Professional Geographer*, 55(1):56-69. Doi: 10.1111/0033-0124.01052

Rogers, A., J. Little and J. Raymer (2010). The indirect estimation of migration. Methods for dealing with irregular, inadequate, and missing data. Dordrecht: Springer. Doi: 10.1007/978-90-481-8915-1

Roy, J., and J. Flood. (1992). Interregional migration modelling via entropy and information theory. *Geographical Analysis* 2416–34. Doi: 10.1111/j.1538-4632.1992.tb00250.x

Samuelson, P.A. (1938). A note on the pure theory of consumer's behaviour, *Economica*, 5(17):61-71. Doi: 10.2307/2548836

Schelling, T. C. (1971). Dynamic models of segregation. *Journal of Mathematical Sociology*, 1(2):143–186. Doi : 10.1080/0022250X.1971. 9989794.

Schelling, T.C. (2006). Micromotives and macrobehavior. Revised edition. New York: Norton (original edition 1978).

Sen, A. (1999). Development as freedom. New York: Anchor Books.

Sen, A. and T.E. Smith (1995). Gravity models of spatial interaction behavior. Berlin: Springer. Doi: 10.ro07/978-3-642.-79880-1

Shannon, C.E. (1948). A mathematical theory of communication. *Bell System Technical Journal,* 27:379-423 and 623-656. Available at http://people.math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Sherlock, C., P. Fearnhead and G.O. Roberts (2010). The random walk Metropolis: Linking theory and practice through a case study. *Statistical Science*, 25(2):172-190. Doi: 10.1214/10-STS327

Shin, J.K. and H. Sayama (2014). Theoretical investigation on the Schelling's critical neigborhood demand. *Communications in Nonlinear Science and Numerical Simulation*, 19:1417-1423. Doi: 10.1016/j.cnsns.2013.08.038

Simon, M. (2019). The effects of immigration policy on migration systems. PhD Thesis, University College London. Available at https://discovery.ucl.ac.uk/id/eprint/10069429/1/MS_Thesis_Final.pdf

Simon, M., C. Schwartz, D. Hudson and S.D. Johnson (2018). A data-driven computational model on the effects of immigration policies. PNAS (Proceedings of the National Academy of Sciences), 115(34): E7914–E7923. Doi: 10.1073/pnas.1800373115

Sims, C.A. (2003). Implications of rational inattention. *Journal of Monetary Economics*, 50:665-690. Doi: 10.1016/S0304-3932(03)00029-1

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges (Distribution functions with n dimensions and their margins). *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 299-331. (English translation: https://www.quantsummaries.com/Hamati_Copula%20in%201959.pdf)

Smith, J.Q. (2010). Bayesian decision analysis: Principles and practice. Cambridge: Cambridge University Press. ISBN: 978-0-521-76454-4

Snickars, F. and J.W. Weibull (1977). A minimum information principle : Theory and practice. *Regional Science and Urban Economics*, 7(1-2):137-168. Doi: 10.1016/0166-0462(77)90021-7

Stephan, F.F. (1942). An iterative method for adjusting sample frequency tables when expected marginal totals are known. *The Annals of Mathematical Statistics*, 13(2):166-178. DOI:10.1214/aoms/1177731604.

Thober, J., N. Schwarz a ndK. Hermans (2018). Agent-based modeling of environment-migration linkages: a review. *Ecology and Society*, 23(2):41(pp. 1-34). Doi: 10.5751/ES-10200-230241

Tjaden,J., D. Auer and F. Laczko (2019). Linking migration intentions with flows: Evidence and potential use. *International Migration*, 57(1):36-57. Doi: 10.1111/imig.12502

Train, K. E. (2009). Discrete Choice Methods with Simulation. 2nd Edition. Cambridge, UK: Cambridge University Press.Doi : 10.1017/CBO9780511805271 Available at https://eml.berkeley.edu/books/choice2.html

United Nations (1998). Recommendations on statistics of international migration. Revision 1. Statistical Papers Series M, No. 58, Rev. 1. New York: United Nations, Department of Economic and Social Affairs, Statistics Division. https://documents-dds-ny.un.org/doc/UNDOC/GEN/N98/163/30/PDF/N9816330.pdf?OpenElement

United Nations (2019). International migration 2019. Highlights. New York: United Nations Department of Economic and Social Affairs, Population Division. Document ST/ESA/SER.A/439. Available at https://www.un.org/development/desa/pd/sites/www.un.org.development.desa.pd/files/files/documents/2020/Jan/un_2019_internationalmigration_highlights.pdf

United Nations (2020). International migration 2020. Highlights. New York: United Nations Department of Economic and Social Affairs, Population Division. Document ST/ESA/SER.A/452. Available at https://www.un.org/en/desa/international-migration-2020-highlights

United Nations (2021). Migration statistics. Report of the Secretary-General. Statistical Commission, 52nd session. March 2021, Document E/CN.3/2021/11 (15 December 2020). New York: United Nations. https://digitallibrary.un.org/record/3897154?ln=en and https://unstats.un.org/unsd/statcom/52nd-session/documents/2021-11-MigrationStats-E.pdf (all documents: https://unstats.un.org/unsd/statcom/52nd-session/documents/).

van Wissen, L. and R. Jennissen (2008). A simple method for inferring substitution and generation from gross flows: asylum seekers in Europe. In: J. Raymer and F. Willekens eds. International migration in Europe. Data, models and estimates. Chichester: Wiley, pp. 235-251. ISBN: 978-0-470-03233-6

Warnke, T., O. Reinhardt, A. Klabunde, F. Willekens and A.M. Uhrmacher (2017). Modelling and simulation decision processes of linked lives: An approach based on concurrent processes and stochastic race. *Population Studies*, 71(Supplement 1):S69-S83.

Weidlich, W. and G. Haag (1988). Interregional migration: Dynamic theory and comparative analysis. Berlin/Heidelberg: Springer.

Willekens, F. (1977). The recovery of detailed migration patterns from aggregate data: an entropy maximizing approach. Research Memorandum RM-77-58, Laxenburg, Austria: International

Institute for Applied Systems Analysis. Available at https://www.researchgate.net/publication/254812305_The_Recovery_of_Detailed_Migration_Patterns_from_Aggregate_Data_An_Entropy_Maximizing_Approach

Willekens, F. (1979). Optimal migration policies. An analytical approach. *Regional Science and Urban Economics*, 9:345-367. Doi: 10.1016/0166-0462(79)90004-8

Willekens, F. (1982). Multidimensional population analysis with incomplete data. In: K.C. land and A. Rogers eds. Multidimensional mathematical demography. New York: Academic Press, pp. 43-111.

Willekens, F. (1983). Log-linear modeling of spatial interaction. *Papers of the Regional Science Association* 52: 187–205. Doi: https://doi.org/10.1111/j.1435-5597.1983. tb01658.x.

Willekens, F. (1994). Monitoring international migration lows in Europe. Towards a statistical data base combining data from different sources. *European Journal of Population* 10(1): 1–42. Doi: https://doi.org/10.1007/BF01268210.

Willekens, F. (1999). Modeling approaches to the indirect estimation of migration flows: From entropy to EM. *Mathematical Population Studies* 7(3): 239–278. Doi: https:// doi.org/10.1080/08898489909525459.

Willekens, F. (2011). La microsimulation dans les projections de population (Microsimulation in population projections). *Cahiers Québécois de Démographie*. 40(2) :267-297. DOI : 10.7202/1011542ar. Available at http://id.erudit.org/iderudit/1011542ar

Willekens, F. (2016). Migration Flows: Measurement, Analysis and Modeling. In International Handbook of Migration and Population Distribution, edited by M. White. International Handbooks of Population 6, 225–241. Dordrecht: Springer. Doi: https:// doi.org/10.1007/978-94-017-7282-2_11.

Willekens, F. (2017). The Decision to Emigrate: A Simulation Model Based on the Theory of Planned Behaviour. In A. Grow and J. van Bavel eds. Agent-based Modelling in Population Studies: Concepts, Methods and Applications, 257–299. Dordrecht: Springer. Doi: https://doi.org/10.1007/978-3-319-32283-4_10.

Willekens, F. (2021). The emigration decision process: Foundations for modelling. Deliverable 2.3. QuantMig. Available at http://www.quantmig.eu/res/files/QuantMig%20D2.3%20V1.1.pdf

Willekens, F., D. Massey, J. Raymer, and C. Beauchemin. (2016). International migration under the microscope. *Science* 352(6288): 897–899. Doi: https://doi.org/10.1126/ science.aaf6545.

Willekens, F., S. Zinn, and M. Leuchter. (2017). Emigration rates from sample surveys. An application to Senegal. *Demography* 54(6): 2159–2179. Doi: https://doi.org/ 10.1007/s13524-017-0622-y.

Wilson, A. G. (1970). Entropy in urban and regional modelling. London: Pion.

Wiśniowski, A. (2017). Combining labour force survey data to estimate migration flows: the case of migration from Poland to the UK. *Journal of the Royal Statistical Society A*, 180(1):185-202. Doi: 10.1111/rssa.12189

Wolpert, J. (1965). Behavioral aspects of the decision to migrate. *Papers of the Regional Science Association*, 15:159-169. DOI: 10.1007/BF01947871

Yaméogo, B.F., P. Gasteneau, P. Hancack and P.-O. Vandanjon (2021). Comparing methods for generating a two-layered synthetic population. *Transportation Research Record*, 2675(1):136-147. Doi: 10.1177/0361198120964734

Yasuda, N. (1975). The random walk model of human migration. *Theoretical Population Biology*, 7:156-167. Doi: 10.1016/0040-5809(75)90011-8

Ye, P., X. Hu, Y. Yuan and F.-Y. Wang (2017). Population synthesis based on joint distribution inference without disaggregate samples. Journal of a-Artificial Societies and Social Simulation. 20(4):16. Doi: 10.18564/jasss.3533

Ye, P. and X. Wang (2018). Population synthesis using discrete copulas. *Proc. IEEE Int. Conf. Intell. Transport. Syst. (ITSC)*, 2018, pp. 479–484 In: 21st International Conference on Intelligent Transportation Systems (ITSC). Doi: 10.1109/ITSC.2018.8570021

Zaloznik, M. (2011). Iterative proportional fitting. Theoretical synthesis and practical limitations. PhD Thesis, University of Liverpool. Available at http://www.researchgate.net/publication/262258986

Zinn, S. (2014). Package MicSim. Performing continuous-time microsimulation. Published on CRAN. https://cran.r-project.org/web/packages/MicSim/index.html

Zwanzig, R. (1983). From classical dynamics to continuous time random walks. *Journal of Simulation Physics*, 30(2):255-262. Doi: 10.1007/BF01012300

Annex A Entropy maximization of univariate distribution

Consider a discrete random variable X with an unknown probability mass function $p_X(x)$. What is the most probable probability mass function? If all we know about the distribution is that $\sum_x p_X(x) = 1$, entropy maximization tells us that the most probable distribution is the uniform distribution. This example is simple but has all the ingredients of any entropy maximization. The problem is to find the most probable distribution $p_X(x)$ that meets the condition imposed on the distribution and does not introduce any assumption on the type or shape of the distribution.

The most probable values of $p_X(x)$ are obtained by solving the following mathematical programming problem:

$$\text{maximize } \mathrm{H}[p_X(x)] = -\sum_{x\epsilon R} p_X(x)\, ln\, p_X(x)$$

$$subject\ to \sum_{x\epsilon R} p_X(x) = 1 \qquad (A.1)$$

Notice that $\mathrm{H}[p_X(x)]$ is a positive value because the logarithm of a probability is negative. The common approach is to replace this constrained mathematical programming problem by an

equivalent problem without the constraint. In mathematical programming, the original problem with the constraint is known as the *primal* problem and the equivalent problem without the constraint as the *dual* problem. The primal and the dual represent two different perspectives on the same problem. The dual is obtained by adding the constraint to the objective function. That requires an indicator of the change in the objective function if the constraint is not met or relaxed. It is the price to pay by not meeting the constraint. The indicator is the Lagrange multiplier. The unconstrained function to be maximized is the *Lagrangian function*:

$$L(p_X(x), \lambda) = -\sum_{x \epsilon R} p_X(x)\, ln\, p_X(x) + \lambda \left[\sum_{x \epsilon R} p_X(x) - 1\right] \qquad (A.2)$$

where λ is the Lagrange multiplier. The values of $p_X(x)$ and the Lagrange multiplier that maximize $L(p_X(x), \lambda)$ must be determined. Since the function reaches a maximum when the slope is zero, the first-order conditions are[12]

$$\frac{\partial L}{\partial p_X(x)} = -(\ln p_X(x) + 1) + \lambda = 0$$

$$\frac{\partial L}{\partial \lambda} = \sum_{x \epsilon R} p_X(x) - 1$$

Hence $p_X(x) = exp[\lambda - 1]$ with $\sum_x p_X(x) = 1$. The probability is the same for all possible values of the random variable X. Since these probabilities must add to one, the probabilities are $p_X(x) = 1/r$. Hence the maximum entropy distribution with a single constraint that all probabilities add to one is the uniform distribution. In other words, if all we know about the probability distribution to be estimated is that the probabilities must add to one, then the distribution that maximizes the entropy and satisfies the principle of indifference is the uniform distribution. The distribution is determined entirely by the information constraint and considers no other knowledge. The distribution is maximally uncertain (information content is the lowest possible). The entropy of the distribution is $\ln r$, with r the cardinality of X. The Lagrange multiplier is $\lambda = 1 + \ln\left(\frac{1}{r}\right) = 1 - \ln r$.

Annex B Data preparation

Introduction

The main source of data used in this paper is the United Nations. The study of international migration suffers from a lack of comparable data. To arrive at comparable data and harmonized data collection procedures, the United Nations (1998) published recommendations on statistics of international migration. Many countries followed the recommendations, but several did not. The United Nations is currently in the process of updating the guidelines. As part of that endeavour, an Expert Group on Migration Statistics was established. The Expert Group developed an overarching conceptual framework on statistics on international migration and mobility and proposed a set of definitions to guide de collection and harmonization of migration statistics (United Nations, 2021). In Section B.2 the framework is briefly described. It is helpful to describe the data types used in this paper.

The study of international migration suffers from the lack of comparable data. The United Nations recommendations on statistics of international migration aim at comparable and harmonized

[12] The derivative of x ln(ax) is ln(ax)+1

statistics (1998, 2021). The framework proposed by the United Nations, revised in 2021, is used to guide the documentation of the data used in this paper.

The term 'countries' is used to denote countries and independent territories (see Section B.3). International organizations, such as the United Nations and the European Commission, use different codes to denote countries. The list of units varies between organizations and sometimes also between publications of the same organization. Section B.3 briefly discusses peculiarities of lists and codes used in international migration statistics.

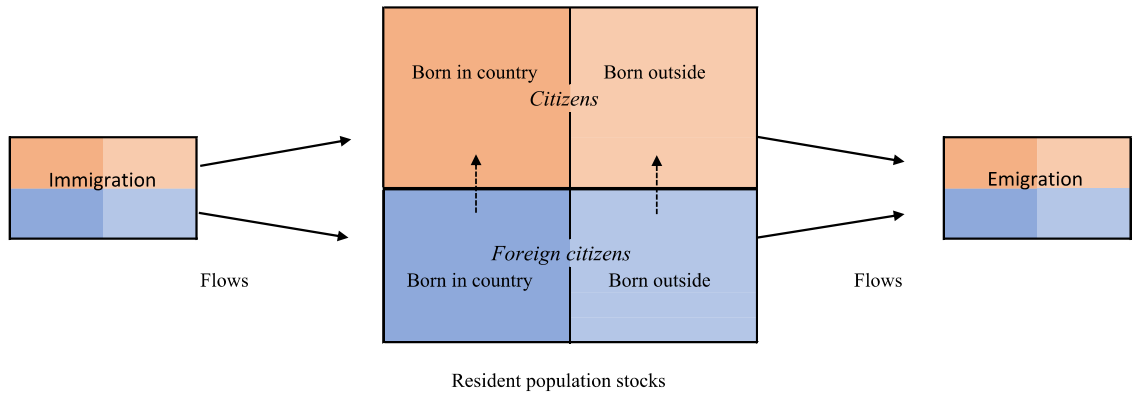
The population data used in this paper are presented in Section B.4. They are from the *2019 Revision* of *World Population Prospects* (United Nations, 2019[13]). The data are used to estimate the number of individuals who never left their country or territory of birth and the estimate the number of persons with current residence the equal to the residence five years ago. These *stayers* are omitted in the United Nations' migrant stock data. Section B.5 covers the two sources of the migration data. The first is the United Nations (2019)' estimates of migrant stocks by country or territory of current residence and country of birth (or nationality). The second source provides data on five-year bilateral flows between countries, which are referred to as 'recent migration'. The estimates are produced by Abel and Cohen (2019). To illustrate the methods presented in this paper, countries are aggregated into six regions. The data for the system of six regions are presented in Section B.6.

Framework on international migration and mobility statistics

The framework distinguishes between resident population and temporary (non-resident) population. International migration is all border crossings related to changes in the resident population. A border crossing that does not change the resident population is international temporary mobility. Migration is therefore a change in the resident population. One aim of the framework is to resolve the current misalignment between flows of international migration and stocks of immigrant populations. "Migration flows generally include all persons immigrating or emigrating to or from a country. Immigrant populations are measured primarily using information on country of birth, without reference to the duration of stay and excluding those who have migrated previously but returned to their country of birth. Without this alignment, there are critical gaps in the evidence base for national policymakers" (United Nations, 2021, p. 7). The framework should be valid irrespective of the concept of resident population used. The UN recommends to define a resident as a person who stays in a country for a duration that exceeds a minimum threshold (ideally 6 or 12 months). According to the United Nations, the resident population should be disaggregated by both birthplace and by citizenship. The conceptual framework is shown in Figure B.1.

[13] Update in United Nations (2020).

Figure B.1 Conceptual framework on international migration and the coherence between flows and stocks



Source: United Nations (2021, p. 10) https://unstats.un.org/unsd/statcom/52nd-session/documents/2021-11-MigrationStats-E.pdf

Changes in stock of migrants are determined by migration flows. Flows measure changes in country of residence. A change in place of residence is an event. However, data on events are frequently not comparable because differences in time frame. Ideally, events that can occur at any time are recorded in continuous time. Migration is such an event. Continuous observation and reporting is not feasible. It is approximated in a national registration system, e.g. a population register. The common approach is to measure changes of residence by comparing places of residence at two points in time. Such data are referred to as *transition data* to distinguish them from event data, also known as *movement data* (Rees and Willekens, 1986). The event concept emphasizes the occurrence. A migration is an event. The transition concept emphasizes the person who has experienced an event. The person's place of residence at one point in time differs from the place of residence at a previous point in time. A transition is measured by comparing a personal attribute, e.g. place of residence, at two points in time. A transition is a consequence of the occurrence of an event. Transition data are collected in censuses and sample surveys, in which persons are asked about their current place of residence and the place of residence at a previous point in time. The previous point in time may be fixed for every respondent, e.g. five years prior to the census, or vary between respondents, e.g. date of birth. The indirect measurement of migration by comparing the current place of residence and the place of birth gives information on *lifetime migration* (at least one migration in a lifetime up to the current age). The indirect measurement of migration by comparing the current place of residence and the place of residence one or five years ago gives information on *recent migration*. The number of transitions recorded or estimated depend on the length of the interval considered. In addition, transition-based definitions of migration undercount the total number of migrations (moves) during the period considered. Statistical techniques have been developed to harmonize data collecting using different time frames. The key idea is to estimate a latent true relocation rate. The approach is proposed by Nowok (2010), who adopts a probabilistic framework and uses simulation because analytical solutions do generally not exist (see also Nowok and Willekens, 2011). Raymer et al. (2013) and Wiśniowski (2017) adopt a Bayesian framework to estimate the latent migration rate. Del Fava et al. (2019) combine the two approaches and extend it using more data.

Countries, areas or territories, and groups of countries

The data sources used in this paper differ slightly in the list of countries and territories for which observations and estimates are presented. In addition, no unique country codes exist. To be able to merge the data from the different sources, a single set of unique codes was prepared. The preparation of the list is documented in this section. The unique country codes form the basis for flexible grouping of countries.

The world consists of 252 countries, independent territories and special areas of geographical interest. 193 are member states of the United Nations and two countries have observer status (Holy See and State of Palestine). Western Sahara is listed by the UN as a "non-self-governing body". Kosovo is not a member of the UN since the UN has not recognized Kosovo as an independent state.

The United Nations' *Standard Country or Area Codes for Statistical Use* identifies countries and territories by three-digit numerical codes. The codes were initially published by the UN Statistics Division as Series M, No. 49 and are now commonly referred to as the M49 standard. The codes are available online (UNSD-Methodology.csv, available at https://unstats.un.org/unsd/methodology/m49/overview/). A total of 249 countries and territories are included in the list. The information provided by the UNSD for each country/territory is shown in Table B.1.

In addition, countries have a three-digit alphabetical code (letter code) and a two-digit alphabetical code assigned by the International Organization for Standardization (ISO): ISO 3166-1 alpha-3 and ISO 3166-1 alpha-2. The latest version is available online at http://www.iso.org/iso/home/standards/country_codes.htm. The official names of countries are available at the UNTERM website at http://unterm.un.org.

No unique list of countries and no unique country codes exist. Organizations use different lists of countries and country codes. For instance, the European Commission and Eurostat generally uses ISO 3166-1 alpha-2 codes with two exceptions: EL is used to represent Greece, not GRC and GR that is used by the UN, and UK is used to represent the United Kingdom, not GBR and GB used by the UN. The codes used by Eurostat are online at https://rdrr.io/cran/eurostat/man/harmonize_country_code.html

Some data published by the United Nations are not available at country level, but are available for groups of countries. The United Nations groups countries into five geopolitical regional groups: African Group, Asia and Pacific Group, Eastern European Group, Latin American and Caribbean Group (GRULAC), and Western European and Others Group (WEOG). Cyprus, an EU member state, is neither a member of WEOG or the Eastern European Group. Due to its geographical location and close ties with Russia, Cyprus decided to remain neutral between the two European Groups and thus is a member of the Asia and the Pacific Group. The United Nations uses numeric location codes for areas of the world. In these codes, the United Nations makes a distinction between UN code and location code. Location codes are used in the world population prospects and the migrant stock data. The UN code for the world is 001, while the location code is 900. The UN code for Europe is 150 and the location code is 908. The country/area codes in M49 are the UN codes. The codes for the world and groups of countries are the UN codes (001 for the World, 002 for Africa, etc.). The location codes of *countries* and independent *territories* are the same as the UN codes used in M49. These codes are less than 900. Location codes of 900 and higher are reserved for the world and groups of countries.

Table B.1 Standard country and area codes for statistical use (M49)
[1] "Global.Code" [2] "Global.Name" [3] "Region.Code" [4] "Region.Name" [5] "Sub.region.Code" [6] "Sub.region.Name" [7] "Intermediate.Region.Code" [8] "Intermediate.Region.Name" [9] "Country.or.Area" [10] "M49.Code" (changed into M49) [11] "ISO.alpha2.Code" (changed into ISO.alpha2) [12] "ISO.alpha3.Code" (changed into ISO.alpha3) [13] "Least.Developed.Countries..LDC." (changed into LDC) [14] "Land.Locked.Developing.Countries..LLDC." (changed into LLDC) [15] "Small.Island.Developing.States..SIDS." (changed into SIDS) [16] "Developed...Developing.Countries" (changed into Developed_Developing)
Source: UNSD https://unstats.un.org/unsd/methodology/m49/overview/

The list of countries changes periodically due to the formation of new countries. For example, Sudan (SDN) included South Sudan, which was founded in 2011 and received the country code (SSD). In that year, the numeric code of Sudan was also changed. Codes are dynamic too. For changes in ISO codes, see https://www.iso.org/iso-3166-country-codes.html and https://en.wikipedia.org/wiki/ISO_3166-1_alpha-3. Statistics Canada prepared an overview of current and historical countries and areas of interest (see https://www.statcan.gc.ca/en/subjects/standard/sccai/2011/scountry-desc). In 2006, Serbia and Montenegro (SCG) was divided into Serbia (SRB) and Montenegro (MNE). In 2008, Serbia was divided into Serbia (SRB) and Kosovo (XKO).

County names change too. In 1989, Burma (BUR) changed into Myanmar (MMR) and in 2009 back to Burma. The UNSD code is Myanmar (MMR).

Table B.2 shows the list of countries and areas and their identifications. The identifications form the basis for flexible grouping countries. In the list, the United Kingdom of Great Britain and Northern Ireland (official name) is replaced by United Kingdom. United Kingdom includes Scotland, Wales, England and Northern Ireland, and excludes Isle of Man (IMN), the Channel Islands (CHI) [Guernsey (GGY) and Jersey (JEY); until 2011, Sark (XSQ) was part of Guernsey] and British Overseas Territories. Some countries/areas are included in the list, but not included in the United Nations list of 249 countries/areas. They are:

- China, Taiwan Province of China (TWN)
- Channel Islands (CHI)
- Sudan before 2011 (SUD)
- Serbia and Montenegro (SCG)

The countries and territories are added because some UN data include these countries/areas. This brings the total to more than 252 countries/areas.

A function was developed in R to produce different groups of these countries/areas. One aggregation is into six regions. That aggregation is used in this paper.

Population data

Official population estimates and projections are prepared by the Population Division of the Department of Economic and Social Affairs of the United Nations Secretariat. For this paper data are obtained from the *2019 Revision* of *World Population Prospects* (wpp2019). Population estimates are given for 201 countries and territories, and 48 groups of countries/areas (total 249) from 1950 to 2020 (5-year intervals). The areas are (United Nations, 2019):

- World (1)
- Regions (5): Africa, Americas, Asia, Europe, Oceania
- Subregions (18): "Northern Africa" ,"Sub-Saharan Africa", "Latin America and the Caribbean", "Northern America", "Central Asia", "Eastern Asia", "South-eastern Asia", "Southern Asia", "Western Asia", "Eastern Europe", "Northern Europe" , "Southern Europe", "Western Europe", "Australia and New Zealand", "Melanesia", "Micronesia", "Polynesia"
- Intermediate regions (9): "Eastern Africa", "Middle Africa", "Southern Africa" "Western Africa" "Caribbean", "Central America" "South America" "Channel Islands"

The data are available online at https://population.un.org/wpp/ [14]. They can also be accessed via the data query https://population.un.org/wpp/dataquery/. The data are also available as the wpp2019 package in the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/web/packages/wpp2019/index.html). In wpp2019, Channel Islands is included, but not the territories that form the Channel Islands. Taiwan (TWN) is included as "China, Taiwan Province of China". Notice that the population for mid-2020 is a projected population.

Migrant data

a. Migrant stock data

The UN migrant stock data presents estimates of international migrants by age, sex, country and area of birth and country and area of current residence for the years 1990, 1995, 2000, 2005, 2010, 2015 and 2020. An update was published in 2020. The estimates are prepared by the UN Population Division and based on official statistics on the foreign-born population (if country of birth is given). If the country or territory of birth is unknown, the UN uses nationality. If neither country/area of birth or nationality is known, then migrants are counted in the resident population with place of birth and nationality unknown. Refugee data are included if available. The 2020 migrant stock estimates cover 232 countries and territories, and 45 groups of countries. The names and composition of geographical areas follow those presented in "Standard country or area codes

[14] See also the `wppExplorer`, which Allows to interactively explore data from the World Population Prospects, contained in packages wpp2019, wpp2017, wpp2015, wpp2012 and wpp2010. It is based on the shiny package. https://cran.r-project.org/web/packages/wppExplorer/wppExplorer.pdf

for statistical use" (ST/ESA/STAT/SER.M/49/Rev.3), available at http://unstats.un.org/unsd/methods/m49/m49.htm

The 2019 data are available online at https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp. In December 2020, an update of the 2019 data was published (IMS2020). The update includes migrant stock data for mid-2020.

The United Nations estimated the number of international migrants (stock) worldwide at 281 million in mid-2020, up from 248 in 2015. For a considerable number of immigrants, the place of birth is not known. The UN estimates whether the origin is in the South or the North and provides the estimates under area labels "Other South" and "Other North". The world population and the sizes of the migrant stocks in the world are shown in Table B.3. M_stockUN is the total migrant stock and M_stock is the migrant stock excluding "Other South" and "Other North". The latter figures are used in this paper.

For this paper, the total international migrant stock data of 2019 are used.

b. Migrant transition data

Abel and Cohen (2019) estimate migration between 202 countries/areas. The estimates are for five five-year periods from mid-year (July 1) 1990 to mid-year (June 30) 2015. The estimates were updated in 2021, using the most recent International Migrant Stock (IMS2020) data made available by the United Nations, and the most recent World Population Prospects (WPP2019). In addition, the time series of data was extended to 2015-20. The most recent estimates, including the period 2015-20, are used in this paper.

The countries/areas include in the estimates differ slightly from those listed by the UN. For instance, the authors include the Channel Islands, but not the separate territories the UN considers (Jersey, Guernsey and Sark, which is in fact part of Guernsey).

Abel and Cohen review and validate six methods recently proposed to estimate recent bilateral migration flows from migrant stock data published by the UN. Bilateral migration data are presented in square contingency tables. The off-diagonal entries contain the migration stocks or flows, depending on the estimates used. The diagonal entries contain the number of native-born residents (stock data) or number of individuals who did not migrate (stayers) or migrated within a country or area. In the assessment, intra-country migration is zero. Validation was done by comparing the estimated produces with reported data where possible. Recall that transition data measure recent changes in residence for persons present at the end and at the beginning of the interval. Since birth and deaths during the interval affect the number of transitions, Abel and Cohen introduce estimates of births and deaths.

Abel and Cohen (2019) apply multiple methods for estimating migration flows from stocks to produce estimates of the bilateral international migration flows between all pairs of 200 countries for five five-year periods from mid-year (July 1) 1990 to mid-year (June 30) 2015, using the set of bilateral migration stocks published by the United Nations. Migration estimates for 2015-19 are not included in the 2019 paper. The authors found that the two methods that use a demographic accounting approach perform consistently better than the four other estimation approaches. The accounting method frames changes in migrant stocks as residuals in a global demographic account (Abel, 2013). Migration flows are estimated to match increases or decreases in the reported

bilateral stocks of migrants, and births and deaths during the period. Of the 232 countries in the UN migrant stock database only 200 had complete estimates of other demographic measures such as births and deaths. They are considered by Abel and Cohen. All of the 32 excluded countries had populations below 100,000. The reported data used to validate the estimation methods are immigration, emigration and net migration flows and population data. All these data are reported by the United Nations Population Division (UNPD). The UNPD adjusted data to include available refugee statistics. The flow data collected and reported by the UN are for single years, whereas Abel and Cohen estimate flows over a 5-year period. To approximate 5-year flows, they multiplied the average annual flows by five. The UNPD and the authors' net migration data represent migrant transitions (based on migrants' location at the beginning and end of the 5-year interval) rather than the number of moves during the interval. Hence, migration to third countries and return moves during an interval are not accounted for.

The validation exercise revealed that estimates based on Azose and Raftery (2019) method agree better with the available data than the estimates produced by the other five methods. In 2021, the authors updated the estimates of directional migration flows using the newly published International Migrant Stock (IMS2020) data inputs by the United Nations, the most recent World Population Prospects (WPP2019), while accounting for new countries added recently. The estimates are available on Figshare (https://figshare.com/articles/dataset/Bilateral_international_migration_flow_estimates_for_200_countries_1990-1995_to_2010-2015_/7731233?backTo=/collections/Bilateral_international_migration_flow_estimates_for_200_countries/4470464). They are referred to as Abel and Cohen (2021).

Table B.3 shows the sizes of migrant stocks and 5-year transitions for different years/periods. The population for 2020 is the projected population (wpp2019).

Table B.3 Population, migrant stocks and 5-year transitions (millions)

	Population	M_stockUN	M_stock	namperiods	5yr-transitions
1990	5306	153	144	1990-95	69.593
1995	5722	161	153	1995-00	67.011
2000	6121	173	166	2000-05	75.851
2005	6518	191	183	2005-10	87.091
2010	6933	220	212	2010-15	93.049
2015	7355	248	237	2015-20	95.864
2020	7770	281	260		

Source: United Nations wpp2019 (columns 1-4) and Abel and Cohen (2021) (columns 5-6)

System of six regions

The United Nations groups countries into different types of regions (see Table B.2). For the presentation of the model, the world is viewed as a system of six regions. The regions are:

a. EU+ (EU_EFTA_UK): (32 countries/areas): 27 member states of the European Union, plus the 4 countries of European Free Trade Association (EFTA) (Iceland, Liechtenstein, Norway and Switzerland), and the United Kingdom. Countries are identified by their three-digit numerical codes:
    i. EU: "040","056","100","191","196","203", "208","233","246","250","276","300","348","372","380","428","440","442","470","528","616","620","642","703","705","724","752"

ii. EFTA: "352","578","438","756"
iii. UK: "826"

b. USCan: United States of America ("840") and Canada ("124")
c. Latin America and the Caribbean: the subregion of countries defined by the UNSD (see Table B.2) (52 countries/areas).
d. Africa: the region of countries defined by the UNSD (see Table 1)
e. Asia: the region of countries defined by the UNSD (see Table B.2), except Cyprus. Cyprus is included in EU+. The region consists of 50 countries/areas.
f. Rest of the world (56 countries/areas): all countries/areas included in list produced by the UNSD (Table B.2), that are not included in (a)-(e).

Note that Channel Islands (Jersey, Guernsey) (British Crown Dependencies of the French coast of Normandy), which are included in the Abel-Cohen study, are included in the category "Rest of the world". Isle of Man (self-governing British Crown Dependency) and Svalbard and Jan Mayen Islands (jurisdiction of Norway),which are included in the UNSD list of countries/areas, are not considered by Abel and Cohen.

Table B.4 shows the population of the six regions. They are obtained by aggregation of the data provided in the *2019 Revision* of *World Population Prospects.*

Table B.5 shows migrant stock data. The off-diagonal elements show the number of people (in thousand) by region of birth (origin) and region of residence (destination) in the year indicated (year 2000 and 2020). The diagonal entries indicate the number of people (in thousand) born in a region and residing in the same region but in a different country of that region in the year indicated. The intra-regional lifetime migration as a proportion of the total lifetime migration in a period is 48.05 percent in 2000 and 50.81 percent in 2020.

Table B.4 Population in six regions of the world, 1990-2020 (millions)

	Year						
Region	1990	1995	2000	2005	2010	2015	2020
EU_EFTA_UK	487.810	494.428	497.705	506.818	515.894	522.383	527.554
USCan	279.662	294.328	312.299	327.158	343.159	356.905	368.745
LatAm	442.574	482.735	521.537	557.172	590.992	623.553	653.561
Africa	630.343	717.264	810.978	916.149	1039.299	1182.433	1340.592
Asia	3204.854	3470.876	3718.354	3954.253	4185.294	4408.757	4616.031
Rest	260.619	262.289	259.666	256.601	257.912	261.103	263.366
Sum	5305.862	5721.920	6120.540	6518.151	6932.549	7355.133	7769.849

Source: United Nations Population Division, wpp2019

Table B.5 Immigrant stock in 2020 by region of birth (millions)

Region	Region of residence in 2020						
Of birth	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	23.952	5.750	1.376	0.864	2.006	3.781	37.730
USCan	1.065	1.095	1.231	0.067	0.509	0.241	4.207
LatAm	4.898	25.911	8.107	0.036	0.424	0.205	39.581
Africa	10.576	2.871	0.038	21.209	4.572	0.577	39.843
Asia	14.348	18.465	0.404	1.222	66.683	11.542	112.664
Rest	7.625	1.837	0.041	0.069	5.129	10.832	25.532
Sum	62.463	55.928	11.197	23.468	79.324	27.177	259.557

The number of people who, at a given point in time, reside in the region of birth consists of two components. The first is the number of people who, in 2020, reside in their country/area of birth (stayers) and the second is the number of people who migrated to another country/region within the region. The second component, i.e. the number of intra-regional international migrants, is shown in the diagonal of Table B.5. The total number of individuals born in their region of residence in 2020 (stayers) is obtained by subtracting, for each region, the number of individuals living in a country other than their country of birth from the population in the region in 2020. Table B.6 shows the population in 2020 by region of residence and region of birth. A diagonal entry shows the number of residents born in their country of residence or in a different country in the same region. An off-diagonal entry shows the numbers of individuals in a given region, born in a different region.

Table B.6 Population by region of residence in 2020 and by region of birth (millions)

	Destination						
Origin	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	489.043	5.750	1.376	0.864	2.006	3.781	502.821
USCan	1.065	313.912	1.231	0.067	0.509	0.241	317.024
LatAm	4.898	25.911	650.471	0.036	0.424	0.205	681.945
Africa	10.576	2.871	0.038	1338.334	4.572	0.577	1356.967
Asia	14.348	18.465	0.404	1.222	4603.390	11.542	4649.371
Rest	7.625	1.837	0.041	0.069	5.129	247.021	261.721
Sum	527.554	368.745	653.561	1340.592	4616.031	263.366	7769.849

Table B.7 presents a summary of the estimates. The first column is the population in mid-2020. The second column shows the number of residents in a region born in the *country* of birth (stayers). The third column is the number of residents in a region born in another country of that region. The fourth column is the number of residents in a region born in that region. It is the sum of stayers and intra-regional international migrants. The next column shows the number of immigrants, that is residents of a region born in another country in that region or in another region. The next column shows the number of emigrants, that is individuals born in a region (row label) currently residing in a different country in that region or another region. The final column shows the net migration transitions.

Table B.7 Population by region of residence in 2020 and region of birth. Summary table (millions)

	Variable						
Region	Pop	Stayers	Intra	Stayers+intra	Im	Ex	Net
EU+	527.554	465.091	23.952	489.043	503.601	478.868	24.733
USCan	368.745	312.816	1.095	313.912	367.650	315.928	51.721
LatAm	653.561	642.364	8.107	650.471	645.454	673.838	-28.383
Africa	1340.592	1317.125	21.209	1338.334	1319.383	1335.758	-16.375
Asia	4616.031	4536.707	66.683	4603.390	4549.347	4582.688	-33.340
Rest	263.366	236.189	10.832	247.021	252.534	250.889	1.645
Sum	7769.849	7510.292	131.879	7642.171	7637.970	7637.970	0.000

Table B.8 shows the Abel-Cohen estimates of the number of recent migration flows (transitions) within and between the six regions of the world during the period 2015-20. The entries in a column show the number of residents in a region at the end of the interval (2020) by region of residence at the beginning of the interval (2015). The diagonal entries represent international migration transitions within each of the six regions. The estimated total number of migration transitions in the world between 2015 and 2020 is 95.9 million. The addition of stayers to the intra-regional migrants gives, for each region, the number of people in the region in 2020 that was also in the

region in 2015. Several of these persons may not have stayed in the region on a continuous basis. They may have left the region after mid-2015 and returned before mid-2020. The population figures are shown in Table B.9.

A summary of recent flow estimates is shown in Table B.10.

Recall that recent changes in residence are estimated for persons present at the end and at the beginning of the interval. Abel and Cohen added estimates of births and deaths. Migration (transition) probabilities are computed the numbers of transitions are divided by the population at the beginning of the interval (1995 for the 1995-00 interval and 2015 for the 2015-19 interval). The population in 1995 and 2015 by region of residence is shown in Table B.4. The estimation procedure introduces a small error due to the added births and deaths. The difference is considered sufficiently small to disregard the effect of births and deaths. A further justification for this simplification is that, in this paper, the data are used mainly to illustrate the multiregional model with individual preferences and immigration quota.

Table B.8 Immigrant population in 2020 by region of residence in 2015 (million)

	Destination						
Origin	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	9.440	1.283	0.905	1.329	2.018	1.718	16.693
USCan	1.385	0.338	2.920	0.331	2.102	0.434	7.510
LatAm	1.569	4.912	6.281	0.014	0.129	0.079	12.984
Africa	2.709	1.128	0.021	8.455	0.978	0.133	13.424
Asia	5.734	5.490	0.215	0.893	22.381	3.471	38.185
Rest	1.843	0.343	0.030	0.084	1.902	2.867	7.069
Sum	22.679	13.494	10.370	11.107	29.511	8.702	95.864

Table B.9 Population in 2020 by region of residence in 2015 (millions)

	Destination						
Origin	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	514.315	1.283	0.905	1.329	2.018	1.718	521.567
USCan	1.385	355.589	2.920	0.331	2.102	0.434	362.761
LatAm	1.569	4.912	649.472	0.014	0.129	0.079	656.174
Africa	2.709	1.128	0.021	1337.940	0.978	0.133	1342.908
Asia	5.734	5.490	0.215	0.893	4608.901	3.471	4624.705
Rest	1.843	0.343	0.030	0.084	1.902	257.531	261.733
Sum	527.554	368.745	653.561	1340.592	4616.031	263.366	7769.849

Table B.10 Population by region of residence in 2020 and region of residence in 2015. Summary table (millions)

	Variable						
Region	Pop	Stayers	Intra	Stayers+intra	Im	Ex	Net
EU+	527.554	504.874	9.440	514.315	518.113	512.127	5.987
USCan	368.745	355.251	0.338	355.589	368.407	362.423	5.984
LatAm	653.561	643.191	6.281	649.472	647.281	649.894	-2.613
Africa	1340.592	1329.485	8.455	1337.940	1332.137	1334.453	-2.316
Asia	4616.031	4586.520	22.381	4608.901	4593.650	4602.324	-8.674
Rest	263.366	254.664	2.867	257.531	260.499	258.866	1.633
Sum	7769.849	7673.985	49.762	7723.747	7720.087	7720.087	0.000

The migration transitions in the period 1995-2000 are used to approximate location preferences. Table B.11 shows the population in 2000 by region of residence in 1995. Of the 495 million people

in the EU+ in 1995, 489 million are in the EU+ in 2000, which is 98.75 percent. In Africa, the probability of staying is 99.54 percent. Less than 0.5 percent of the population of Africa had a residence outside of Africa in 2000. Most who left, went to Europe, 0.27 percent of the population of Africa and 58.02 percent of those who left. The revealed location preferences are shown in Table B.12.

Table B.11 Population in 2000 by region of residence in 1995 (million)

	Destination						
Origin	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	489.252	2.052	0.495	0.756	1.466	1.425	495.446
USCan	0.804	298.729	1.734	0.096	1.140	0.153	302.656
LatAm	0.676	6.214	519.123	0.006	0.141	0.025	526.185
Africa	2.174	0.721	0.022	809.694	0.712	0.118	813.441
Asia	2.663	4.118	0.133	0.394	3713.307	3.767	3724.382
Rest	2.136	0.466	0.030	0.032	1.588	254.179	258.431
Sum	497.705	312.299	521.537	810.978	3718.354	259.666	6120.540

Table B.12 Location preferences revealed by the 1995-2000 migration flow

	Destination						
Origin	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	0.987500	0.004141	0.000999	0.001525	0.002958	0.002877	1
USCan	0.002656	0.987027	0.005730	0.000316	0.003766	0.000505	1
LatAm	0.001285	0.011809	0.986579	0.000012	0.000268	0.000047	1
Africa	0.002673	0.000886	0.000027	0.995393	0.000876	0.000145	1
Asia	0.000715	0.001106	0.000036	0.000106	0.997026	0.001011	1
Rest	0.008265	0.001804	0.000115	0.000123	0.006145	0.983547	1

Annex C Create virtual population

In the absence of individual data, the virtual population is created by sampling contingency tables (DeSalvo and Zhao, 2020; Kayibi et al., 2018). The aim is to reconstruct a population that is consistent with the aggregate figures tabulated in contingency tables. In principle, the statistical analysis of virtual populations is not different from the statistical analysis of real populations, except for the variances. The only difference between a virtual population and a real population is that the virtual population is obtained by sampling theoretical probability distributions with parameters estimated from empirical data, whereas real sample populations are obtained by sampling real populations.

A virtual population is created from the tabulated data on population in 2020 by region of residence in that year and region of birth (Table B.6). In 2019, 272 million people or 3.5 percent of the world population resided in a country other than their country of birth (United Nations, 2019). In 2020 the migrant stock increased to 281 million or 3.6 percent of the world population (United Nations, 2020). In the system of six regions, 1.6 percent resided in a region other than their region of birth. Hence half of the international migrants (50.8 percent) reside in the region of birth. The proportion migrating to another region of the world is even less if *recent* changes of residence (2015-20) are considered, instead of lifetime migration data. Abel and Cohen (2019) found that, in 2020, 95.86 million people resided in a country other than their country of birth, which is 1.2 percent of the world population (Table B.8). The proportion that resides in 2020 in a

region other than their region of residence in 2015 is 0.59 percent. Hence 51.9 percent of the international migrants in the period 2015-20 moved to another country in their region.

Excluding stayers from the study of migration is a bad practice, as many migration scholars have recently testified. That applies in particular when the decision to migrate or stay is the subject of study since the decision to stay is as much an expression of agency as the decision to leave the country. In this paper, stayers are included in the analysis. Since 96 percent of the world population resides in the country of birth and even more in the region of birth, a large sample of the world population is needed to obtain a sufficient number of migrants by region of origin and region of destination. For that reason, a random sample of one million people or 0.125 per thousand of the world population is drawn from the available tabulated data.

The individuals in the virtual population are given an identification number (ID), a region of birth, a region of current (2020) residence, and a region of residence in 2015. The procedure used ensures that the virtual population (microsystem) is fully consistent with the tabulated data derived from the country-specific estimates published by the UN (2019) and Abel and Cohen (2019, 2021) (macrosystem). Each individual is assigned a location preference. It is the preferred region of residence based on the preferences revealed by the migrant flow in the period 1995-2000. Table C.1 shows the first records of the person data structure. The first individual is born in Latin America, currently resides in Latin America and prefers to stay in the region. Individual 7, born in Africa, is currently living in EU+, but prefers the USA or Canada. The last column is the identification number of the region of preference. The procedure used to produce these simulated data is described in this Annex. Additional personal attributes, if available and relevant for the analysis, could be easily added using a similar procedure.

Table C.1 Person data structure

	ID	birth	IDc19	IDc15	preference	pref
1	1	LatAm	LatAm	LatAm	LatAm	3
2	2	Asia	Asia	Asia	Asia	5
3	3	USCan	USCan	USCan	USCan	2
4	4	Rest	Asia	Asia	Asia	5
5	5	Asia	Asia	Asia	Asia	5
6	6	Asia	Asia	Asia	Asia	5
7	7	Africa	EU+	EU+	USCan	2
8	8	Africa	Asia	Asia	EU+	1

The Annex consists of three sections. The first describes the allocation of region of birth and region of residence in 2020. The second covers the allocation or region of residence in 2015. The third adds location preferences.

a. Region of birth and region of residence in 2020

Three methods are considered: simple multinomial sampling and two versions of stratified sampling. The first involves simple multinomial sampling from an empirical probability distribution; namely, the joint distribution of region of birth and region of residence in 2020 (Table B.6). The 36 parameters of the multinomial distribution are obtained by dividing the entries in Table B.6 by the overall total (world population). The parameter p_{ij} of the multinomial distribution is the probability that an individual, selected at random from all individuals in the population, is born in region i and resides in region j in 2020. Each individual is assigned a region of birth and a region of residence such that the joint distribution of region of residence and region of birth in the

virtual population is equal to the distribution of regions of birth observed in the 2020 population. Although the empirical region-of-residence-by-region-of-birth distribution is maintained, the entries in Table B.6 are not reproduced exactly (scaled to the size of the virtual population). The reason is the random variation introduced by the sampling, i.e. sample variation.

The second method, stratified sampling, removes the sample variation. The sample population in each combination of region of birth and region of residence in 2020, i.e. in the strata, is obtained by multiplying the number of individuals in the virtual population (one million) by the probabilities computed from Table B.6. The cell entries are not subject to sample variation. The entries are real values, not integers. Integer values are obtained by rounding. To ensure that the rounding errors do not cause the sum of the cell counts to deviate from the total size of the virtual population, the largest cell count is adjusted. The result is shown in Table C.3. An alternative (third) method, stratified and sequential random sampling, produces integer values and ensures that the sample population has exactly the same joint distribution of region of birth and region of residence in 2020 as the observed population. It starts by selecting 62,941 individuals at random from all individuals in the virtual population. They are assigned EU+ as region of birth and EU+ as region of residence in 2020. Next, 740 individuals are randomly selected from the individuals who did not yet get assigned a region of birth and a region of residence in 2020. They are assigned EU+ as region of birth and USCan as region of residence in 2020. The procedure is continued until all individuals have received a region of birth and a region of residence in 2020. Notice that all individuals have equal probabilities to be assigned to a given region of birth and a given region of residence.

Table C.2 Population, by region of residence in 2020

	Population (million)	Percentage	Sample
EU+	527.554	6.79	67898
USCan	368.745	4.75	47458
LatAm	653.561	8.41	84115
Africa	1340.592	17.25	172538
Asia	4616.031	59.41	594095
Rest	263.366	3.39	33896
Sum	7769.849	100.00	1000000

Table C.3 Sample population by region of residence in 2020 and region of birth

	Region of residence in 2020						
Birth	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	62941	740	177	111	258	487	64714
USCan	137	40401	158	9	66	31	40802
LatAm	630	3335	83717	5	55	26	87768
Africa	1361	369	5	172247	588	74	174644
Asia	1847	2376	52	157	592472	1485	598389
Rest	981	236	5	9	660	31792	33683
Sum	67897	47457	84114	172538	594099	33895	1000000

b. Region of residence in 2015

The members of the virtual population are assigned a region of residence in 2015 by stratified and sequential random sampling. For each stratum the sample size is given by the cell entries of Table B.9, scaled to a total equal to the size of the virtual population. The result is shown in Table C.4. The figures are fully consistent with the figures in Table B.9. For example, 66193=1000000*514.315/7769.849. The last two figures are from Table B.9.

Note that the region of residence in 2015 assigned to an individual depends on the region of residence in 2020, but is independent of the individual's region of birth, implying that the relocation in the period 2015-19 is independent of the region of birth. This independence assumption is not realistic, but is made because no three-way classification of individuals is available on region of residence in 2020 by region of residence in 2015 and region of birth. Some countries collect, in censuses and surveys, data from their residents on country of birth and country of residence five years ago. That data could be used to introduce dependencies of migration transition flows on region of birth. Alternatively, it might be possible to extend the estimation method used by Abel and Cohen (2019, 2021) to produce region of birth-specific estimates of recent migration.

c. Location preferences by region of residence in 2015

In this paper, the location preferences are the preferences revealed by the migration flows in the period from 1995-2000, estimated by Abel and Cohen (2019, 2021). Location preferences depend on the region of residence only and are not influenced by other factors. The distribution of location preferences among the residents of a given region is a multinomial distribution: $\{p_{i1}^0, p_{i2}^0, p_{i3}^0, \dots, p_{ir}^0\}$. An element p_{ij}^0 gives the probability that an individual in region i prefers to live in region j. It is determined by the spatial distribution in 2000 of the residents of region i in 1995. Table B.12 in Annex B shows the location preferences. The table is referred to as the matrix of location preferences or preference matrix. Consider the population of Africa in 1995. Of the population of Africa in 1995 the large majority (99.5 percent) are in Africa five years later, 2.7 per thousand reside in EU+, 0.9 per thousand in USACan, etc.. Notice that the revealed preferences account for the immigration restrictions (quota) that existed in the period 1995-2000.

Stratified and sequential sampling is used to assign a preferred region of residence to individuals. Recall that the distribution of location preferences depends on the region of residence at the beginning of the baseline interval, in this case 2015. The procedure consists of two steps. In the first step, the number of residents of i with a preference for region j is obtained by multiplying the number of residents in i by the probability that a resident of i prefers region j. The expected number of people in region i preferring j is $p_{ij}^0\, n_{i+}$. In the second step, $p_{ij}^0\, n_{i+}$ individuals are selected at random from the residents of i who have not been assigned a preferred region of residence yet. The procedure ensures that the number of residents in i who prefer j is equal to the expected number obtained in the first step. The number of individuals in the sample by region of residence at the beginning of the interval (2015) and preferred region of residence are shown in Table C.5. The figures are fully consistent with the figures in Table 3.1. For example, the number of individuals in EU+ in 2015 that prefers the live in EU+ is 66287=1000000*515.047/7769.849. The last two figures are from Table 3.1.

Individual data on region of birth, regions of residence in 2015 and 2020, and location preference are stored in a person data structure (data frame or person file) with one record per person. This approach to data storage has a big advantage; namely, that results of the simulation may easily be added to the data frame.

Table C.4 Sample population by region of residence in 2020 and region residence in 2015

	R2020						
R2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66193	165	116	171	260	221	67126
USCan	178	45764	376	43	271	56	46688
LatAm	202	632	83587	2	17	10	84450
Africa	349	145	3	172196	126	17	172836
Asia	738	707	28	115	593180	447	595215
Rest	237	44	4	11	245	33144	33685
Sum	67897	47457	84114	172538	594099	33895	1000000

Table C.5 Sample population by region of residence at t and preferred region of residence

	Preferred region of residence						
Region2015	EU+	USCan	LatAm	Africa	Asia	Rest	Sum
EU+	66287	278	67	102	199	193	67126
USCan	124	46082	268	15	176	24	46688
LatAm	108	997	83317	1	23	4	84450
Africa	462	153	5	172040	151	25	172836
Asia	426	658	21	63	593445	602	595215
Rest	278	61	4	4	207	33131	33685
Sum	67685	48230	83681	172225	594200	33978	1000000