Journal of Computational Social Science
https://doi.org/10.1007/s42001-023-00226-7

RESEARCH ARTICLE

Modeling economic migration on a global scale

Eva Dziadula[1] · John O'Hare[1] · Carl Colglazier[2] · Marie C. Clay[1] · Paul Brenner[1]

Received: 13 February 2023 / Accepted: 12 September 2023

Abstract
We introduce a global-scale migration model centered on neoclassical economic migration theory and leveraging Python and Jupyter as the base modeling platform. Our goals focus on improving social scientists' understanding of migration and their access to visually and computationally robust infrastructure. This will enhance a scientist's capability to model complex macro-scale global effects and lay the groundwork for multi-scale models where countries, regions and individuals interact at differing timescales and per differing governing equations. Economic theory describes an agent's migration decision as utility maximizing. The agent weighs the expected increase in utility associated with migration against the costs of moving to that destination. These costs include not only the explicit monetary costs of travel and visas, but also the implicit costs such as leaving family behind, political barriers to entry, the difficulty in learning a new language, and the unfamiliarity of a new culture, among others. In our model, any destination country in which an agent would have greater earnings (minus migration costs) than in the origin country is considered and agents maximize their expected earnings. Multiple public data sets from United Nations, International Monetary Fund, and World Bank are used to provide suitable initialization values for the model. Our Global Open Simulation (GOS) software has an open license and the data analyzed during the current study are available in the GOS public Github repository (https://github.com/crcresearch/GOS).

Keywords Human migration · Population change · Jupyter · Python

✉ Paul Brenner
paul.r.brenner@nd.edu
https://scholar.google.com/citations?user=zCsBL7sAAAAJhl=en

[1] University of Notre Dame, Notre Dame, IN, USA

[2] Northwestern University, Evanston, Il, USA

Published online: 09 October 2023

Introduction

Motivation

According to the United Nations, there were 281 million international migrants in 2020, an increase of 62 percent since 2000 with nearly two-thirds of all international migrants residing in high-income countries [40]. Human migration throughout the world affects global economies, regional stability and the lives of many individuals, be they migrants themselves or members of communities affected by migrants. For example, the United States is one of the top destinations for economic migrants worldwide, and more than 13 percent of the population is foreign born [42]. Furthermore, the Department of Homeland Security estimates that over 12 million of these migrants are undocumented [2]. 2017 and 2018 witnessed a major political struggle to replace the DACA (Deferred Action for Childhood Arrivals) child immigrant policy [43], establish new immigration enforcement measures and revise immigration policy to favor merit in visa allocation over family cohesiveness. Not all migration, however, is for economic gains. Forced migration resulting from refugee flight can result in numerous social problems for both the recipient country and the country of origin (not to mention the tremendous strain on the refugees themselves). According to the United Nations High Commissioner for Refugees, there were 89.3 million people in 2021 who were forcibly displaced from their homes—the highest measured since the second world war. This issue has been highlighted with distress in areas of the world such as the Middle East and more recently Ukraine. For instance, in 2015 large refugee flows from Syria into Europe created a humanitarian crisis at the borders of European countries [19]. This overwhelmed European states with additional unfunded infrastructure demands for shelter, food, education, work, etc.

Attempts to handle large-scale migration have been met with mixed success. European Union leaders recognized the growing demand for aid and provided resources for refugees and migrants traveling from Africa and the Middle East to Greece or Italy [32]. There are also some success stories of Syrian migrants able to relocate and adjust to a new life in Germany [9]. In this light, we recognize the major challenges to understanding the multi-scale complex systems that influence regional and global population change. At a global scale, countries can be modeled as agents, each with individual characteristics that influence immigration and emigration relative to all other countries. At the same time, internal dynamics of regional “state" level agents can influence the aggregate country characteristics over time. For example some states might economically benefit from immigration but other states more vocally object to immigrants’ social impact on their communities. Lastly, at the scale of the individual one could mix the influence of government and regional policies with the individual’s personal traits and nuclear family influences.

Developing powerful and accessible global migration simulation tools will increase our capacity to understand and predict why and when migrations will occur and where outbound migrants are most likely to go when they leave their places of origin. We include agent-centric migration motivations and external mobility limiting factors such as relative financial status, transit routes, established immigration laws, etc. Our model accuracy is coupled with the accuracy of governing data sets (which measure population, historical migration, economic productivity, etc.) garnered from world organizations and published literature. Furthermore, the mechanisms employed for resolving and interpreting conflicting, disparate, and missing data are key to improving the model's utility. Finally, the simulation tool must be powerful, enable data visualization/analytics, and be accessible to a broad range of social scientists. These factors motivate our selection of the open source Python and Jupyter web-based research and development interface.

Related work

Social problems are difficult to solve via computational models due to the complexity of their nature, and often the root cause is systematic [44]. International migration specifically is a deeply complex phenomena that continues to be studied by expert cross-disciplinary teams such as Castles, Haas and Miller [8]. Further, there is historically demonstrated unpredictability of social systems and errors in computational modeling can lead to wildly different outcomes. In comparison, statistical analysis of large and broad sets of social data has shown increasing promise at deriving human interaction patterns and correlations [14]. We view first principles simulation and data analytic techniques (such as machine learning) as complementary components to the development of more accurate predictive models and social understanding. In this work, we focus our modeling efforts based on complex, yet determinant, statistical relationships which are then perturbed via random variation of country policies [6, 23, 30].

A few Agent-Based Models (ABM)s have attempted to simulate human migration, albeit not on a global scale. One country-scale ABM by Williams [45] presented an agent-based model that simulated the migration from neighborhood to neighborhood in Nepal. Williams used empirical data from the Chitwan Valley Family Study to initialize the behavior and characteristics of each agent. While this model was specific to rural migration in Nepal, the ideas of the model can be extended to other countries with different variable characteristics. They implemented a stochastic model where the probability that an agent will emigrate was $Log(P/(1-P))$ based on a series of weighted factors such as age, sex, and economic status multiplied by their respective coefficient weights from regression equations of these characteristics. For each individual, if the probability an agent will migrate that month is greater than a randomly generated number, the agent would make the decision to move accordingly. An application of ABM to migration in Vietnam by Nguyen et al. [27] incorporates the theory of planned behavior to identify the intention and contributing factors to migration as it further widens the rural–urban inequality in the region. They model migration decisions for different demographic groups. Searle and van Vuuren [34] apply ABM to forced migration using the conflict in Syria. They highlight the difficulty of obtaining reliable data pertaining to migration shocks that are not of economic nature and also the importance of building models that can assist in predicting the associated migration flows which is where our model comes in.

When studying the global human population (over eight billion people) the sheer size makes the system a challenge to approach. Simulating a global system where each individual agent exhibits its own behavior and has its own data values increases costs in computational time and memory needed to run a simulation. In terms of scale, over the past decade, a number of researchers have demonstrated that large population agent-based simulations are viable. In 2011, Parker et. al [29] used GSAM to implement a graph-based model of disease propagation among 6.75 billion people using 32 CPU cores and 256GB of memory. Our prior work has demonstrated the ability to scale to 7 billion agents for simple reference social models and individual migration decisions [4, 18]. In this work, we focus on a macro-scale migration model with each country as an agent with unique properties which govern the statistical outcome.

Migration model

Economic foundation

Economic opportunities are one of the primary forces behind the decision to migrate. Mayda [25] showed that with each thousand-dollar increase in GDP per capita in the host country, the percentage of immigrants increases by half a percentage point. Income in the country of origin is also important as it determines whether individuals have a reason to go and also whether they can afford to go. At low levels of GDP per capita, individuals may not be able to leave, and at high levels of income, they may no longer want to [11]. The size of ethnic enclaves in the destination (the presence of other migrants, especially from the same country of origin or speaking the same language) also increases the likelihood of migration to that destination [3]. Furthermore, the ease of the actual move, whether it is the geographic distance [17], proximity in language or culture [1], or the generosity of immigration policy [28], also impact the migration choice.

Neoclassical migration theory describes an agent's migration decision as a utility-maximizing decision. The agent will weigh the benefits of moving against the cost of moving to that destination. In the absence of a comprehensive understanding of each agent's preferences, we can assume that the primary incentive for economic migrants is their expected income and can use the expected wages as a proxy in the migration decision[24]. More recently, Kennan [20], who also focused on expected earnings as the main determinant of migration, developed a model which allows for many migration choices and a sequence of location decisions, as many migrants either return or migrate again. If the increase in expected income in the destination outweighs the costs of moving there, a rational individual would act on this increase in utility and decide to move to the destination. Any country in which an agent would be happier than in the origin country will be considered a potential destination in our model. If multiple destinations provide the agent with an increase in utility, we would expect the agent to choose destinations with greater net increases.

We use Borjas [7] adaptation of the Roy model which focuses on wealth maximization as a baseline for our model. Agents migrate if expected lifetime earnings in destination $E(W_D)$ minus the cost of migration (C) outweigh the earnings in origin. An agent will migrate if:

$$E(W_D) - C > W_O. \tag{1}$$

This simplified version of the model assumes a discounted present value of lifetime earnings in both locations. We focus on a one-time static approach in this first version of our model and return migration would simply be considered another possible move from one country to another in a future period. We, thus, abstract from determining whether the migration flows are transitory or permanent and their short- or long-run effects on the countries of origin or destination. While these general equilibrium outcomes are interesting, they are beyond the scope of the current model. We simply focus on the flexibility and ability of users to change the model's parameters to simulate hypothetical events to predict the induced flows. Our model does not have the automated capability to model longitudinal migratory effects and, thus, we do not model the resulting dynamic changes. The migration decision is complex. While the decision is ultimately a cost–benefit analysis, the degree of perceived costs and benefits is different for each individual. Leaving family behind may be a small cost for some, while for others it may be so large that they will never migrate regardless of potential earnings in the destination. Moreover, while economic gains are at the forefront of this decision for many, there are instances of forced migration where security becomes the primary concern and the role of income is diminished. The migration decision is then no longer a carefully evaluated calculation of future income gains but rather a swift decision to save their lives and they can be transitory or permanent [26]. Again, our model allows the user to calibrate the importance of these factors and adjust the modeled prediction.

Modeling framework

One of the most reliable predictors of migration is the difference in income potential, often proxied for by GDP per capita and levels of inequality. Both of these are incorporated in our model within the return to skill function (RTS) we constructed. The World Development Indicators [36] provide the income shares of each country by population quintile. We use these data along with each country's GDP and population size to calculate the average GDP per capita in each quintile. Then, we approximate the income distribution by fitting an exponential line of the form Ae^{Bz} to the average per capita incomes of each population quintile calculated from the data. A and B values are country-specific constants allowing for the best fit of the return to skill function. In this framework, z captures 100 percent of the population ordered by income, and income is used as a proxy index for the agent's skill level. Given that wages reflect the marginal product of labor and are, therefore, a function of skill, skill is strongly correlated with income [16]. The income distribution captures the wages of every skill level within the country and can, therefore, be used as a proxy for the distribution of skill. In the Borjas model, the return to skill is related to income inequality in the country. Based on the relative return to skill in the destination and origin, we may observe positive or negative selection in migration. If the return to skill is higher in the destination, then high-skill agents will migrate. If the return to skill is lower in the destination, in other words, there is less income inequality, then low-skill agents will migrate as the lower end of the wage distribution represents a higher absolute wage than in origin. In our RTS function, the corresponding y values are in US dollars and highlight the reward to skill in each country. Figure 1 highlights the different RTS functions for a few selected countries: United States, Mexico, and Syria.

An agent's expected earnings in any potential destination country can be approximated using the information on the agent's skill level, how the destination country rewards skill, and the probability of finding a job there. The former two factors are incorporated in our RTS function and we use the unemployment rate as the lower bound approximation of the difficulty of finding a job in the destination.

Expected earnings for an agent in a destination country D are:

$$E(W_D) = P(Employment_D) \cdot RTS_D(x), \tag{2}$$

where x is the agent's skill level and RTS is the return to skill function estimating the agent's income in country D based on his or her skill level x. In the classic Borjas framework, an agent migrates if the expected earnings in destination, less the cost of migration, are larger than the earnings of the agent in his or her country of origin, where the current earnings are known W_O. However, in our model, to allow for flexibility of estimating the choices of all the possible agents in the origin, we use the expected earnings for the selected skill level and account for the probability

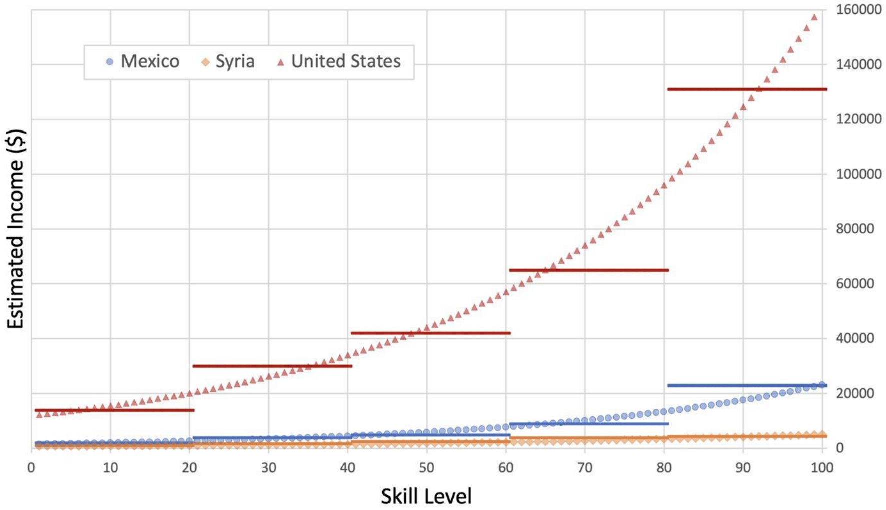

Fig. 1 The estimated RTS curves for the United States, Mexico, and Syria

of having a job in the origin (unemployment rate). Expected earnings for an agent in the origin country O are:

$$E(W_O) = P(Employment_O) \cdot RTS_O(x). \tag{3}$$

Therefore, in our framework, an agent migrates if:

$$P(Employment_D) \cdot RTS_D(x) - C > P(Employment_O) \cdot RTS_O(x). \tag{4}$$

In a theoretically pure model with perfect and dynamically adjusted cost considerations, if multiple country destinations satisfy this inequality, the agent should choose to migrate to the destination country which maximizes the value, conditional on being positive:

$$max[P(Employment_D) \cdot RTS_D(x) - C - P(Employment_O) \cdot RTS_O(x)]. \tag{5}$$

We do not, however, believe that we are modeling all costs perfectly and dynamically a priori. For example we recognize that there are limits to the number of immigrants that even a resilient host country will accept en mass from a single origin [33]. Further, there are stability limits to the number of emigrants from a single origin in terms of the rapid change in cost factors the origin location experiences as population shrinks. Therefore, where there exist multiple economically viable migration destinations we assign individuals a destination not based on the maximum but based on a probability likelihood based directly on the relative percent economic benefit one destination yields over the other.

The cost of migration (C) varies based on the relationship between the origin and destination countries. Some characteristics of the countries act as pull factors toward the destination and others as push factors from the origin. We incorporate both of these into our basic choice as part of expected earnings. For our baseline model of costs, the factors we have chosen to incorporate are the geographic distance, migration history from the country of origin to the destination, whether the countries share a language, and measures of openness to immigration and political freedoms in the destination. A greater physical distance between the origin and destination will be associated with a higher cost of migrating. The explicit costs of travel between two countries will be greater if the countries are farther apart, so both the migration itself and any return trips will be more costly. Furthermore, the agent will be faced with the psychological cost of family and friends being less accessible the farther away they are. In our model, distance (D) is measured as the great circle distance between the average latitude and longitude of each country. A greater distance between countries corresponds to a greater cost of migration.

$$D_{O \leftrightarrow D} = \frac{D \text{ (great circle distance between origin and destination)}}{D_{Max}}. \tag{6}$$

A weaker history of migration from the origin to the destination will also present a greater cost, as the destination will be less familiar to the agent. We measure out-migration (OM) as the current percentage of country O's population residing in country D. Initialized with historical data from the United Nations[13], migration history is tracked within the model as agents move. Greater historical rates of migration from origin to that particular destination correspond to a lower cost of migration.

$$OM_{O\to D} = \frac{\text{migrants from origin in destination}}{\text{population of origin}}. \tag{7}$$

Moreover, the larger the concentration of migrants is in the destination, the more developed the ethnic enclave may be providing more resources and access to information and jobs, and thus acting as a pull factor. We measure ethnic enclave (EE) as the current percentage of country D's population that migrated from country O.

$$EE_{O\leftrightarrow D} = \frac{\text{migrants from origin in destination}}{\text{population of destination}}. \tag{8}$$

We define migration history (MH) as the weighted sum of out-migration (OM) and ethnic enclaves (EE):

$$MH_{O\to D} = \gamma_1(1 - OM_{O\to D}) + \gamma_2(1 - EE_{O\to D}) \tag{9}$$

or

$$MH_{O\to D} = 1 - \gamma_1 OM_{O\to D} - \gamma_2 EE_{O\leftrightarrow D} \quad \text{where} \quad \gamma_1 + \gamma_2 = 1. \tag{10}$$

Smaller differences in the language and culture of the two countries will present a smaller psychological cost (pull factor), as assimilation will be easier for the agent. Agents are assigned proficiency in languages spoken in their origin country. If the language spoken in the destination (L) maps to one of the same top three languages spoken in the origin country [41], the cost is zero. Moving to a country with entirely new languages presents a higher binary migration cost.

$$L_{O\leftrightarrow D} = \begin{cases} 0 & \text{if O and D share a spoken language,} \\ 1 & \text{otherwise.} \end{cases} \tag{11}$$

Finally, greater political barriers (PB) will present both explicit costs in the from of passport and visa fees, as well as implicit costs such as difficulty in the application process and time spent navigating regulations. Political barriers combine passport index scores and freedom index scores [15] for each country. We collect the latest "Welcoming Countries Rank" from www.passportindex.org for every destination country and normalize it into an index. The minimum of the passport index (PI) represents the most welcoming countries (pull factor), whereas the minimum of the Freedom index (FI) correspond to the least politically free countries. We weigh the FI of the origin country (push factor) against the FI of the destination country (pull factor) by subtracting the values and transforming them into an index.

$$PB = \delta_1 \frac{PI_D}{100} + \delta_2\left(1 - \frac{(FI_D - FI_O)}{100}\right). \tag{12}$$

Variations in personal preferences will mean that each agent will face differences in cost even between the same origin and destination. The model in its current form accounts for these differences through a probabilistic decision process based on the magnitude of the expected net increase in earnings. The cost C is the weighted average of the following variables, each of which is normalized to 1 (value / max value):

$$C = \alpha_1 D_{O\leftrightarrow D} + \alpha_2 MH_{O\rightarrow D} + \alpha_3 L_{O\leftrightarrow D} + \alpha_4 PB. \tag{13}$$

For estimating the costs of migration, the model allows the researcher to vary the components as well as the weights associated with them. From the existing literature estimating the determinants of multilateral migration [25], geographic distance appears to be more important than the difficulty of learning a new language (lower magnitude and not significant in some specifications). These studies did not account for political climate or migration history. Several studies focusing on migration to a specific country, often the United States, do include these measures [7]. Further supports the role of geographic distance [10]. Highlight the role of the stock of migrants from the origin present in the destination as well as the large impact of policy, immigration quotas in particular. Therefore, for initial weights in the cost function, based on the existing evidence in the literature, we chose to place a larger emphasis on the geographic distance (α_1=.3), migration history (α_2=.5), and somewhat smaller weights on language (α_3=.1) and political barriers (α_4=.1).

The normalization of the cost variable results in a value between 0 and 1. To scale this value, we introduce an adjustable multiplier β on the cost value. The RTS function for each country returns an estimate of income in US dollars. The role of the Beta value is to convert costs into US dollars to match the RTS income estimates. Beta reflects the maximum cost an emigrant can face while considering migration. In our model, we select a unique Beta value for each country, setting the value equal to the median output of the RTS function of that country. This value was selected in order to reflect the higher relative costs emigrants of higher earning countries would face.

Mathematical model

Migration from an origin to a destination may only occur if:

$$E(W_D) > W_0 + \beta \cdot C. \tag{14}$$

Cost is the normalized sum of each cost variable:

$$\begin{aligned} C = {} & \alpha_1 \frac{D_{O\leftrightarrow D}}{D_{Max}} + \alpha_2 \left(1 - \gamma_1 OM_{O\rightarrow D} - \gamma_2 EE_{O\rightarrow D}\right) \\ & + \alpha_3 L_{O\leftrightarrow D} + \alpha_4 \left(\delta_1 PI_D + \delta_2 \left(1 - \frac{(FI_D - FI_O)}{100} \right) \right). \end{aligned} \tag{15}$$

The RTS function takes the following form, where A and B are country-specific constants and z captures 100 percent of the population ordered by income.

$$RTS = Ae^{Bz}. \tag{16}$$

Expected wages at a given skill level are the product of the probability of employment and the output of the RTS function at that skill level:

$$E(W_D) = P(Employment_D) \cdot RTS_D(x). \tag{17}$$

We use the same equation for the agent's wage at the origin. Though it represents the agent's probability of current wage rather than expected wage:

$$W_O = P(Employment_O) \cdot RTS_O(x). \tag{18}$$

Expanding the migration inequality (1) we get:

$$P(Employment_D) \cdot RTS_D(x) > P(Employment_O) \cdot RTS_0(x) + \beta \cdot Cost. \tag{19}$$

The magnitude of migration from origin (O) to destination (D) can be estimated by the magnitude of the migration inequality:

$$\delta_{O\to D} = P(Employment_D) \cdot RTS_D(x) - (P(Employment_O) \cdot RTS_O(x) + \beta \cdot Cost). \tag{20}$$

If $\delta_{O\to D}$ is positive, the inequality holds, and there is a probability of migration. If $\delta_{O\to D}$ is negative, no migration occurs, and the value can be ignored. As such, any negative $\delta_{O\to D}$ are considered as zero in the subsequent calculations.

Two levels of normalization on the $\delta_{O\to D}$ value occur within the model. The first normalized $\delta_{O\to D}$ value is denoted $\rho_{O\to D}$. This value is the percentage likelihood an outbound migrant from O would move to D relative to the other potential destination countries. This likelihood becomes the percentage of migrants moving from O to D relative to all migrants leaving O when enforced. It is calculated by dividing $\delta_{O\to D}$ by the sum of $\delta_{O\to D}$ values for every destination country.

$$\rho_{O\to D} = \frac{\delta_{O\to D}}{\Sigma\left(\delta_{O\to D_i}\right)}. \tag{21}$$

The second normalization on the $\delta_{O\to D}$ value is used to determine the relative size of the emigrant population leaving each country. We denote this normalized value as μ_O and the country with the largest economic incentive for emigration as country M. μ_O is calculated for each origin country O by dividing the sum of country O's $\delta_{O\to D}$ values by the sum of country M's $\delta_{M\to D}$ values.

$$\mu_O = \frac{\Sigma\left(\delta_{O\to D_i}\right)}{\Sigma\left(\delta_{M\to D_i}\right)}. \tag{22}$$

The percentage of migrants leaving a given origin is determined within our model by the equation:

$$Migrant\,Percent\,Population_{O\to D} = \lambda \cdot \mu_O \cdot \rho_{O\to D}, \tag{23}$$

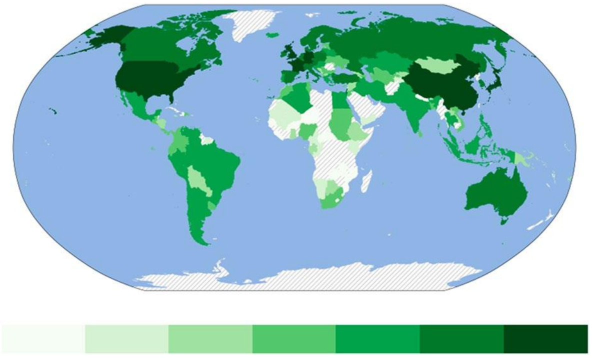

Fig. 2 Magnitude of immigration of skill level 90. Hashed texture indicates incomplete data

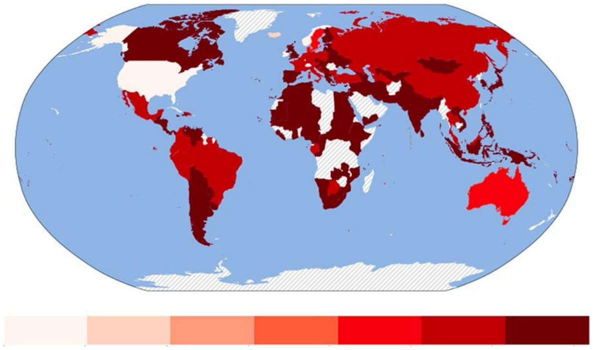

Fig. 3 Magnitude of emigration of skill level 90. Hashed texture indicates incomplete data

where the μ_O value determines the proportion of migrants leaving origin country O. The ρ value determines what share of the migrants leaving country O are going to destination country D. λ is currently set as a global multiplier that prevents more than a set percentage of a country's population from migrating at a given time. We recognize that this is a very rough initial heuristic and hope to determine country-specific values for resiliency and stability, in addition to allowing for large-scale disruptions during war times; for this work, the global value is based on a fairly steady state global economic system.

Results

The graphical results from the global migration model at target skill levels reveal predicted immigration and emigration on a per-country basis as shown in example Figs. 2 and 3. These estimated values are the result of each component of the multifaceted migration decision being weighed in both origin countries as well as in each potential destination. In each country, the cost of migration is weighed against the potential gains of relocation to every possible destination, and some of the dynamics of this decision can be observed in the mapped results from running the model for both high and low-skilled migrants.

In much of the developed world, our results predict large amounts of immigration at both high and low skill levels. Developed nations such as those in Western Europe and the United States have strictly higher RTS curves compared to the majority of the world. This means that the majority of potential migrants would receive higher incomes in these developed destination countries compared to remaining in their origin. In many cases, this positive change in income is large enough to significantly outweigh the cost of migrating. Whenever this is true migration occurs, resulting in the large numbers of immigrants observed. Because the gap in RTS curves exists for the entirety of the curve, it makes sense that we observe similar immigration numbers for both high and low skill levels in these developed nations. The RTS curves in the developed nations are high enough above much of the rest of the world that individuals of any skill will see the benefits of moving outweighing the costs, as modeled. However, for many agents the non-monetary cost of leaving their family behind outweighs the expected monetary gains.

In the rest of the world, predicted immigration tends to scale up with skill level. This can be explained by comparing RTS functions across the world. At the lower ends of the RTS curves, there is not much variation across nations. RTS curves tend to diverge further at greater values, and thus high-skilled agents could stand to gain much more from migration in certain situations than low-skilled migrants. Many of these lower-earning destination countries are attracting more high-skilled immigrants primarily because they offer a low-cost destination for their even lower-earning neighbors, cost permitting.

Predicted emigration from the most developed nations is not as uniform across both country and skill levels compared to immigration to these countries. Excluding parts of Southern Europe, there is very little emigration from these countries at low skill levels. Because of the low variation at this end of the RTS curve, the greatest potential increase in wage will rarely outweigh the costs, and low amounts of emigration are predicted. At higher skill levels, a wider diversity of emigration from these developed countries is predicted. Two factors are contributing to this diversity. The first factor is the greater divergence in RTS curves. These higher gains are more likely to exceed the cost of migration. The result of this factor is high-skilled migrants from already high-earning countries moving to the highest-earning countries such as the United States. The second factor is the wide array of costs associated with the many potential destinations a migrant would be willing to consider. In Europe, even the highest-earning countries like Germany are losing skilled emigrants, whereas in North America, the United States loses relatively few emigrants. German migrants have many more low-cost and high-earning destinations to select from, and in the United States, the only comparable relationship is Canada.

This relation, while most easily observed in Europe and North America, also tends to hold for the rest of the world. In general, the magnitude of emigration increases with the skill level of the migrants, as higher-skill emigrants have more to gain in absolute terms. This is especially true where the cost of migration is higher than normal, for example in island nations such as Japan and Australia. One notable exception in our predictions is Mexico, perhaps capturing the large degree of low-skill migration from Mexico to the United States.

Model validation

Validation procedures

Our validation focuses on the overall quality of the prediction derived from the global migration model. Based on economic theory and existing research, we introduce a number of factors into the composition of cost in the modeling framework and set up the baseline model. The weights for each component in the cost function are as follows:

$$C = 0.3 \cdot D_{O \leftrightarrow D} + 0.5 \cdot MH_{O \rightarrow D} + 0.1 \cdot L_{O \leftrightarrow D} + 0.1 \cdot PB.$$

The model validation is based on the comparison of the model predictions and actual migration data as well as the applicability to different countries. Furthermore, we want to show the reader an alternative specification of the model where we alter the weights and put a larger emphasis on political barriers.

$$C = 0.1 \cdot D_{O \leftrightarrow D} + 0.1 \cdot MH_{O \rightarrow D} + 0.1 \cdot L_{O \leftrightarrow D} + 0.7 \cdot PB.$$

The model can be altered in the code by the user to place a larger weight on certain parameters. We selected eleven countries to show the results of the validation and the alternative model specification with altered weights in the next subsection.

Data for validation

The validation data set is constructed from available World Bank data:

Step 1: Calculate the crude growth rate of the population for each country based on the crude birth rate [37] and crude death rate [38].
Step 2: Multiply the crude growth rate with the population of 2017 to get the net population change only caused by birth and death.
Step 3: Calculate the net population change based on the population of 2018, and subtract the value from Step 2 to derive the net population change caused by migration.

Results analysis

The model and code base allow for variation of all parameters and weights described previously in the paper. While validating the impact of each individual variable is beyond the scope of this paper, we want to provide the reader with a baseline validation of the model using globally recognized and recent historic migration data. To this end, we are using the population data from the World Bank to compare with the predictions of our default model as outlined above, as well as with the predictions of the alternative specification that shifts a larger weight to political barriers in the migration cost.

The blue bars represent the migration estimates calculated using the World Bank data. For Brazil, Finland, Iceland, South Korea, and Spain the estimates are positive and indicate an inflow of migrants, whereas Bangladesh, Indonesia, Mexico, Pakistan, Philippines, and Syria experience an outflow of migrants. The orange middle bar represents the estimated net flows of our baseline model. The direction of migration is consistent and estimates track reasonably well with the World Bank data given the large scope of a global model. The estimated flows are comparable for Iceland and Mexico. The model slightly overestimates migration from Indonesia, Pakistan, and the Philippines. We observe that the model overestimates the inflows to Brazil, Finland, and South Korea and underestimates the inflows to Spain and outflows from Bangladesh and Syria (Fig. 4).

In the alternative specification shown here, we choose to emphasize the role of political barriers in the migration decision and deemphasize the role of migration history and geographic distance (gray bar). The model prediction for migration to Brazil is reduced and is now closer to the World Bank estimates. However, in the case of Indonesia, it results in an even larger overestimation of the outflow of migrants, and an even larger underestimation of migration to Spain. However, in the case of Bangladesh and Syria, the predictions of the alternative model do move closer to the actual flows, which is understandable as the political environment and unrest played a large role. Our model allows the user to change the relative weights of importance on the migration decision. Moreover, users are able to directly specify new values of the underlying parameters and directly override the baseline values. For example in the case of Syria, one could manually change

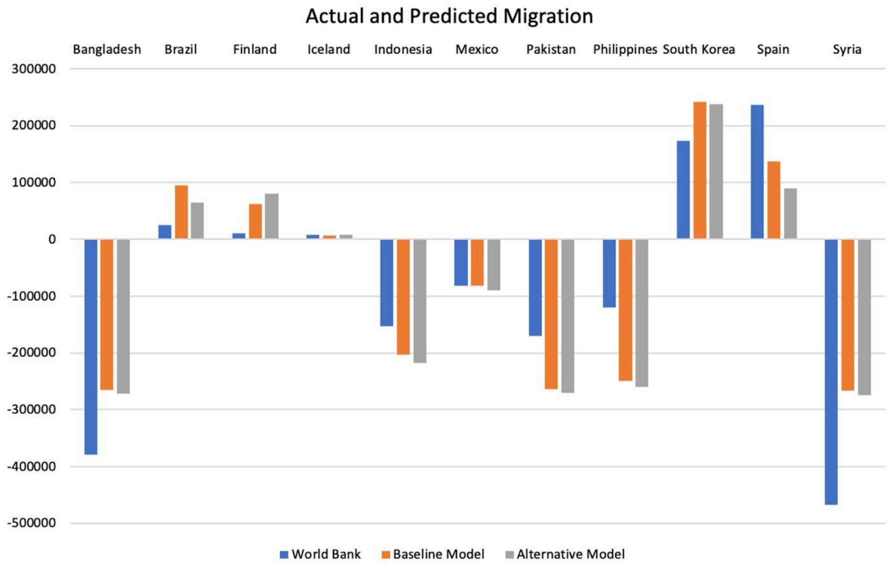

Fig. 4 Validation results

the value of political stability and estimate the predicted flows with the new set of values.

With a comprehensive set of countries with different sets of conditions and circumstances, it is nearly impossible to adapt a simple model that would track perfectly for all. Therefore, users could perform a similar exercise of aligning the model's prediction with real data for the specific country they are interested in, and alter the parameters to match the migration flows. Then, the baseline model would track and the user can experiment with hypothetical shocks to predict migration flows.

Implementation

Language selection

At the University of Notre Dame we have observed significant growth in the use of open source scripting languages such as Python and R which leverage a growing number of performant computational libraries such as SciPy and NumPy. To this end, Python has been selected as the programming language for our social scientists seeking a minor in computing and digital technologies (http://cdt.nd.edu). Given the growing familiarity of Python with our social scientists and its first-class integration with Jupyter, Python is the language of choice for our model. Our team fully recognizes that Python is not as computationally efficient as Fortran or C/C++ but with proper profiling and subsequent use of Python libraries written in C/C++ we have achieved reasonable performance in a language accessible to both the developer and user.

Jupyter platform

The model platform was chosen based on a specific requirement that end user domain scientists need not be required to install any software that does not already come with common desktop/laptop operating systems. Specifically, nothing more than a major internet browser such as Microsoft Edge, Chrome, Safari or FireFox. The corresponding author has observed little success introducing ABMs when undergraduate students and non-technical faculty are required to install various software applications and libraries; due to a host of software/hardware compatibility and security complications. For this reason Jupyter (evolved from IPython) [21, 31] was selected as the user and developer platform. Jupyter allows for browser based access to easily share executable code notebooks with built-in graphical libraries integrated with the underlying programming language of choice (Python, R, C#, Julia, etc.). Further, as we are interested in computationally complex global models the web platform can be deployed on suitable enterprise class computer servers, clusters, and clouds removing the requirement for local system computational resources.

Initialization data: organization and structure

Our global migration model leverages numerous public real-world data sets which enable the scientists to test many realistic scenarios. Specifically, we leveraged:

- A & B RTS Constants Calculated from World Development Indicators [36]
- Country Unemployment Data [41]
- Country Latitude and Longitude [35]
- Country Freedom Index [15]
- Country Passport Index (www.passportindex.org)
- Dominant Country Languages [41]
- Country Populations [39]
- Global Migration History [13]

Gathering, cleaning, and aligning such globally representative datasets, however, was a major challenge. One global organization, the United Nations, recognizes 193 member countries and 2 non-member observers. There is not global consensus on this set with the recognition of certain countries such as Taiwan and Palestine creating significant controversy. Further, the notion of a country might not always be the most relevant basis with the World Bank representing its 189 members plus 28 additional separate economies with populations of over 30,000; while the largest global sports organization in the world FIFA recognizes teams from 211 countries/member associations. There is even less consistency in the naming scheme and uniform reporting of data across of the variables we processed. For this publication, data from 134 countries are reported which represents greater than 94% of the global population.

We leveraged Python Pandas dataframes and Numpy vectors and matrices to create a set of functions to make data processing easier while retaining higher-performance Numpy native data structures. Vectors represent data that maps to a corresponding country; matrices represent data associations from one country to another. For example, the population is represented as a vector because each country has exactly one value. The distance between countries, on the other hand, is best represented by a matrix because there is a unique value for each pair of countries.

Because we were working with country data, we needed a way to standardize how each country was represented in the system. One dataset could say ”United Kingdom" while another could say the “United Kingdom of Great Britain and Northern Ireland"; where both names were representing the same country. To solve this, we leveraged the existing ISO 3166-1 standard for country codes. As they were imported, each data set mapped the name of countries to their corresponding ISO code. To aid in validity and continuity, countries were dropped from the model’s data structures if data for a model variable was missing (for that particular country). For reproducibility, GOS logs each country that is dropped from its store and all of the graphical data visualizations (including those in this paper) represent omitted countries with a diagonal hash texture.

Visualization

Geographical representations of data help accelerate the revelation of patterns within the data, highlighting correlations and connections among dimensions, and providing inspiration about complex data [22]. We, therefore, implemented global visualization capabilities within our platform. The Choropleth Map is the most suitable form for our work. It is a thematic map where areas are shaded in proportion to the quantitative variable. It provides an easy way to visualize how a measurement varies across a geographic area or shows the level of variability within a region. We implemented two methods of visualization, using the Basemap (now CartoPy) and Plotly libraries. The former is more generic and open source but complicated; the latter is interactive, feature rich, and convenient, albeit tied to closed source tools and features. We wrapped them as modules respectively, with necessary functions, and provided APIs for the platform users. Thus far, we have focused more time on the fully open source CartoPy to ensure that all computations and visualizations enabled by the platform do so robustly without closed source component dependencies.

Conclusion and future work

We successfully developed and demonstrated a flexible tool to model global migration yielding a visual representation of the predicted movements of the world's population. Our model provides a unique structure for both data ingest and modeling while allowing users to choose different values of parameters and weights in the specification modeled (without any recompilation of the code). This will provide scientists with the opportunity to try numerous values against historical records for verification and validation as well as providing predictive analytics for future trends. Currently, the model uses data for one time period. In the future, we plan to expand the model by adding data and the model could account for trends in population, GDP, and unemployment, based on observed data to improve the predicted migration flows.

We have made the platform fully open source on GitHub (https://github.com/crcresearch/GOS) with hopes that scientists will collaborate to refine the model(s) with not only recommended variable values but also additional cost factors particular to individual community cultures, patriotism, etc. Two specific areas of work: Parameter Optimization and Multiscale Modeling are planned as part of future work and discussed briefly here.

Model calibration and parameter optimization

We built the model such that weights for each factor influencing migration cost can be tuned to reduce the distance between the model simulation results to the real-world empirical data. This will be a challenging task since the parameters are both multi-dimensional and multi-level. We are currently considering the use of differential evolution algorithms to determine optimal variable values. It is also essential to note that simply calibrating model variables to provide outputs that match empirical data is insufficient as many complex combinations of variable values could yield a data fit but would not yield realistic individual variable values which map to well-understood migration theory and known correlations and causality associated with each variable independently. Thus, we must set fairly tight max/min bounds on as many weight variables as possible prior to leveraging optimization routines to calibrate against empirically known populations. We will also work to have sufficient empirical data so as not to underfit the model (too few data sets). Further, we will not overfit the calibration such that it matches the empirical data perfectly but fails for modestly different new data sets. There are many reasons beyond the scope of the model that influence the migration decision and thus a perfect calibration to the economic factors would not be desirable. That said, investigating which factors lead to over or underestimation of migration flows is another area for future research.

Multiscale modeling via integration with individual centric models

In our current model, the agents' skill levels are chosen at the start of the simulation and remain constant throughout. Thus simulation is most accurate for a small number of waves (iterations) of human migration. While this allows us to model movement at a given skill level, it does not account for the long-term effects of skilled populations leaving their origins for destinations with higher returns to skill, a phenomenon called brain drain. Our results show that as the skill level of study is increased, emigration increases across nearly every country. However, there is not an unlimited supply of skilled people in the world, and populations suffering from brain drain will need to make up for skill loss. To enhance our work, we hope to

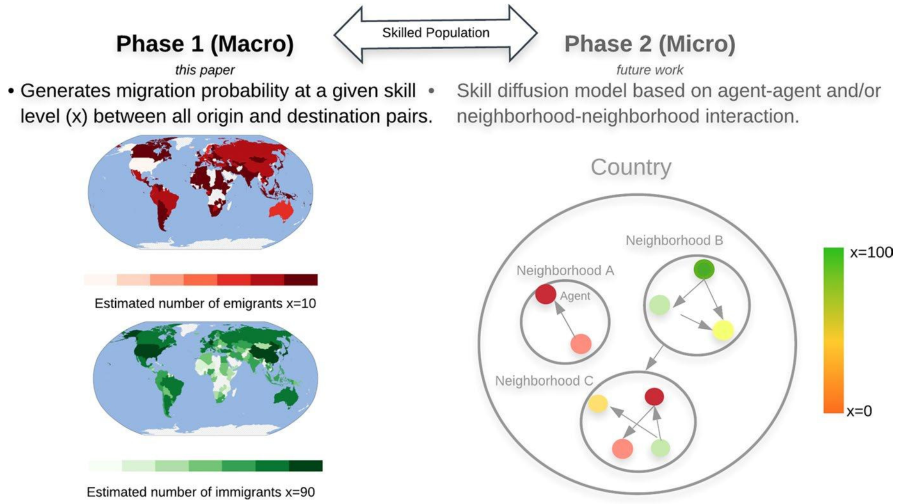
Phase 1 (Macro)
this paper
• Generates migration probability at a given skill level (x) between all origin and destination pairs.
Skilled Population
Phase 2 (Micro)
future work
• Skill diffusion model based on agent-agent and/or neighborhood-neighborhood interaction.
Estimated number of emigrants x=10
Estimated number of immigrants x=90
Country
Neighborhood A
Agent
Neighborhood B
Neighborhood C
x=100
x=0

Fig. 5 Skill diffusion model

make our model capable of running numerous iterations and tracking skilled human movement as populations change over time. Skill will be incorporated as a dynamic variable that changes as interactions occur on an 8 billion agent-scale level. As the model runs and highly skilled agents emigrate, knowledge will regenerate through a process of diffusion that occurs both among and between neighborhoods as demonstrated in Fig. 5. For example, Bomba [5] creates a model of knowledge potential redistribution in social and communicative educational environments, and Cowan [12] studies informal knowledge trading across networks of individuals. Tailoring these models to our work should allow us to capture a more accurate representation of human migration that takes into account the effects of brain drain and the limitations of skill regeneration as populations emigrate.

The authors would like to thank the National Science Foundation for supporting this and related undergraduate research in the computational social sciences at the University of Notre Dame (NSF Award 1560089). They would also like to thank the Notre Dame Center for Research Computing for student and computational resource support. We would also like to thank Fang Guan and Ruyu Zhou who helped with early graphics, parameter optimization and validation work during their time at ND as part of the ND ISURE program. Megan Crawford, Kate Connolly, and William Hurley helped with the validation analysis. Connor Howington helped update code base dependencies and improve portability over subsequent semesters. Finally, we would like to thank the outstanding open source developer community behind the Jupyter platform and Python scientific libraries.

Data availability The Global Open Simulation (GOS) software platform has an open license and the datasets analyzed during the current study are available in a public Github repository (https://github.com/crcresearch/GOS).

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Adsera, A., & Pytlikova, M. (2015). The role of language in shaping international migration. *The Economic Journal, 125*, 586.
2. Baker, B.C.: Immigrant population residing in the united states:january 2014. Office of Immigration Statistics (2017). https://www.dhs.gov/sites/default/files/publications/Unauthorized Immigrant Population Estimates in the US January 2014_1.pdf.
3. Beine, M., Docquier, F., & Özden, Ç. (2011). Diasporas. *Journal of Development Economics, 95*(1), 30–41.
4. Blandin, N., Colglazier, C., O'Hare, J., Brenner, P.: Parallel python for agent-based modeling at a global scale. In: Proceedings of the 2017 International Conference of The Computational Social Science Society of the Americas, CSS 2017, pp. 10:1–10:7. ACM, New York, NY, USA (2017). https://doi.org/10.1145/3145574.3145588.
5. Bomba, A., Nazaruk, M., Kunanets, N., & Pasichnyk, V. (2017). Constructing the diffusion-like model of bicomponent knowledge potential distribution. *International Journal of Computing, 16*, 74–81.

6. Bonabeau, E. (2002). Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences, 99*(suppl 3), 7280–7287.
7. Borjas, G.J.: Self-selection and the earnings of immigrants (1987).
8. Castles, S., De Haas, H., Miller, M.: The age of migration:international population movements in the modern world, 5th edition (2014).
9. Chazan, G.: Syrian refugees in germany: paths diverging. Financial Times (2017). https://www.ft.com/content/304cebc0-08c7-11e7-ac5a-903b21361b43.
10. Clark, X., Hatton, T. J., & Williamson, J. G. (2007). Explaining us immigration, 1971–1998. *The Review of Economics and Statistics, 89*(2), 359–373.
11. Clemens, M. A. (2022). Migration on the rise, a paradigm in decline: The last half-century of global mobility. *AEA Papers and Proceedings, 112*, 257–61. https://doi.org/10.1257/pandp.20221050.
12. Cowan, R., & Jonard, N. (2004). Network structure and the diffusion of knowledge. *Journal of Economic Dynamics and Control, 28*(8), 1557–1575.
13. of Economic, U.D., Social Affairs, P.D..: Trends in international migrant stock: The 2017 revision. (2017).
14. Foster, I., Ghani, R., Jarmin, R.S., Kreuter, F., Lane, J.: Big data and social science: a practical guide to methods and tools. CRC Press (2016).
15. Freedom House: Freedom in the world 2017 (2017). https://freedomhouse.org/report/freedom-world/freedom-world-2017.
16. Griliches, Z., & Mason, W. M. (1972). Education, income, and ability. *Journal of Political Economy, 80*(3), S74–S103.
17. Grogger, J., & Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants. *Journal of Development Economics, 95*(1), 42–57.
18. Howe, A., Brenner, P.: Computational considerations for a global human well-being simulation. The 4th Workshop on Parallel and Distributed Agent-Based Simulations (PADABS) pp. 347 – 355 (2016).
19. Ignatieff, M., Keeley, J., Ribble, B., McCammon, K.: The united states and the european refugee crisis: Standing with allies. Faculty Research Working Paper Series (2016).
20. Kennan, J., & Walker, J. R. (2011). The effect of expected income on individual migration decisions. *Econometrica, 79*(1), 211–251.
21. Kluyver, T., Ragan-Kelley, B., Pérez, F., Granger, B., Bussonnier, M., Frederic, J., Kelley, K., Hamrick, J., Grout, J., Corlay, S., et al.: Jupyter notebooks-a publishing format for reproducible computational workflows. Positioning and Power in Academic Publishing: Players, Agents and Agendas p. 87 (2016).
22. Lee, D.J.: Workshop notes in data visualization and social sciences (2014).
23. Macy, M. W., & Willer, R. (2002). From factors to factors: computational sociology and agent-based modeling. *Annual review of sociology, 28*(1), 143–166.
24. Massey, D. S., Arango, J., Hugo, G., Kouaouci, A., Pellegrino, A., & Taylor, J. E. (1993). Theories of international migration: A review and appraisal. *Population and Development Review, 19*(3), 431–466.
25. Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics, 23*(4), 1249–1274.
26. Naqvi, A., & Monasterolo, I. (2021). Assessing the cascading impacts of natural disasters in a multi-layer behavioral network framework. *Scientific reports, 11*(1), 20146.
27. Nguyen, H. K., Chiong, R., Chica, M., & Middleton, R. H. (2021). Understanding the dynamics of inter-provincial migration in the mekong delta, vietnam: an agent-based modeling study. *Simulation, 97*(4), 267–285.
28. Ortega, F., & Peri, G. (2013). The effect of income and immigration policies on international migration. *Migration Studies, 1*(1), 47–74.
29. Parker, J., Epstein, J.M.: A distributed platform for global-scale agent-based models of disease transmission. ACM Trans Model Comput Simul (2011).
30. Parunak, H.V.D., Savit, R., Riolo, R.L.: Agent-based modeling vs. equation-based modeling: A case study and users' guide. In: International Workshop on Multi-Agent Systems and Agent-Based Simulation, pp. 10–25. Springer (1998)
31. Pérez, F., Granger, B.E.: Ipython: a system for interactive scientific computing. *Computing in Science & Engineering*, *9*(3) (2007).
32. Rumeny, E.: Auditors highlight failings of eu response to migration crisis. Public Finance International (2017).

33. Schiff, M., Ozden, C.: International migration, economic development & policy. The World Bank (2007).
34. Searle, C., & van Vuuren, J. H. (2021). Modelling forced migration: A framework for conflict-induced forced migration modelling according to an agent-based approach. *Computers, Environment and Urban Systems, 85*, 101568.
35. Tamosauskas, T.: Country codes and coordinates (2013). https://gist.github.com/tadast/8827699.
36. The World Bank: Income distribtions (2014). Data retrieved from World Development Indicators, https://data.worldbank.org/data-catalog/world-development-indicators.
37. The World Bank: Birth rate,crude(per 1,000 people) (2017). https://data.worldbank.org/indicator/SP.DYN.CBRT.IN?view=map.
38. The World Bank: Death rate,crude(per 1,000 people) (2017). https://data.worldbank.org/indicator/sp.dyn.cdrt.in.
39. United Nations, D.o.E., Social Affairs, P.D.: World population prospects: The 2017 revision, data booklet. **ST/ESA/SER.A/401** (2017).
40. United Nations, D.o.E., Social Affairs, P.D.: international migration report 2020 (2020).
41. United States Central Intelligence Agency: The world factbook. Retrieved August **20**, 2018 (2018).
42. U.S. Census Bureau: 2014 national projections. U.S. Department of Commerce (2014).
43. Warren, R., & Kerwin, D. (2015). Beyond dapa and daca: Revisiting legislative reform in light of long-term trends in unauthorized immigration to the united states. *J. on Migration & Hum. Sec., 3*, 80.
44. Watts, D.: Computational social science: Exciting progress and future challenges (2016).
45. Williams Nathalie E., M.L.O., Yao, X.: Using survey data for agent-based modeling: design and challenges in a model of armed conflict and population change. In: Agent-Based Modelling in Population Studies, pp. 159–184. Springer International Publishing (2017).

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.