Contextualizing the formation of individual-level migration intentions across the world using machine learning

Alejandra Rodríguez Sánchez Arne Maaß Jasper Tjaden

2024-12-05

Migration intentions are influenced by a wide array of individual-level factors spanning economic, social, demographic, cultural, political, and developmental dimensions. However, the relevance of these drivers is profoundly shaped by the national and regional contexts in which individuals reside. Despite advances in understanding these dynamics, the extent to which the importance of migration predictors varies across countries and regions remains underexplored. Leveraging data from Gallup World Poll (GWP) surveys (2007–2016), we employ supervised and unsupervised machine learning methods, alongside explainable AI techniques, to highlight the role of context in shaping migration intentions. Pooling survey responses across years, we develop an ensemble of ten XGBoost models to predict migration intentions, using most of the core GWP variables as predictors. SHAP values are computed to evaluate feature importance, and dimensionality reduction techniques (PCA and UMAP) are applied to reveal patterns in the salience of predictive factors across countries and regions. Our findings identify a core set of common predictive factors alongside substantial heterogeneity, with regional clustering emerging as a key feature of the formation of migration intentions. These results underscore the importance of country-specific and regional contexts in the predictive strength of migration drivers, suggesting unique geographic dynamics in the formation of migration intentions. By quantifying these variations, our machine learning framework offers new avenues for understanding migration behavior and enhancing the accuracy of migration forecasting models.

Keywords: Migration, Machine Learning, Survey, Explainability AI

Introduction

The regulated and humane management of international migration is a central focus of many governments' migration policies. In a context of increasingly diversifying migration flows, scholars and practitioners have turned to machine learning methods to improve on current policies with the goal of anticipating specific migration flows and be better prepared to attend the incoming population (Clemens et al. 2018). Approaches to forecasting migration often employ a compilation of large scale and real time data sources, as well as adaptive statistical models that consider the changing importance of theoretically and empirically selected migration drivers over time (Carammia, Iacus, and Wilkin 2022; Napierała et al. 2022). Many factors, such as communication and transportation systems, international crises, climate change, demographic characteristics, and labor market challenges, as well as other developments are said to inform migration decision-making, including the formation of migration intentions (Massey et al. 1993). According to recent literature overviews (Aslany et al. 2021; Soto Nishimura and Czaika 2024), migration aspiration determinants can be broadly classified into a small number or subset of dimensions: *demographic, economic, environmental, human development, individual and household resources, politico-institutional, security, socio-cultural, and supranational.* However, in a global examination of mobility patterns, socioeconomic factors were found to be more strongly associated with migration patterns than climatic factors (Niva et al. 2023), showing that not all drivers matter equally across countries.

Changes in any one of these factors, as we see, for example, with unprecedented weather events, the escalation of armed conflicts, or the onset of economic crises, could affect migration flows in some periods, while not in others, further complicating the predictability of migration (Carammia, Iacus, and Wilkin 2022; Migali et al. 2018). However, it is unclear which of these dimensions are more important in predicting migration aspirations across countries, meaning the weight of the evidence for each of these factors is not clear. Moreover, mobility (internal or international) vary greatly across sociodemographic and environmental contexts (Kraemer et al. 2020). One reason why literature reviews find important variation in the salience of specific migration drivers might be indicative of the influence of national or regional context. This could be further indicative of a fundamental inadequacy of migration theories that are not empirically supported across contexts. In other words, the poor predictive ability of migration drivers across countries could be considered an empirical test of the relevance of country-context nuances needed for theoretical development. This high complexity of migration systems makes predictions on the migration behavior of individuals highly uncertain, including the formation of migration intentions, but potentially highly informative to advance migration research.

This uncertainty, however, constitutes an untapped potential for the development of new research questions and methods in the field of migration research. Supervised machine learning models are useful when trying to predict a target variable but their application in some domain, however, can further reveal the limitations of the data and the theoretical models. The failures of a machine learning model could highlight the (un)predictability of certain aspects of human migration, such as migration intentions (Tjaden, Auer, and Laczko 2019). Moreover,

observations that are harder to predict, of which are mispredicted by a machine learning model, could open up new research questions (I. Lundberg et al. 2024), such as which observations are harder to predict, and why. Finally, the application of unsupervised machine learning techniques, such as dimensionality reduction methods and clustering, could further our understanding of the heterogeneity of migration drivers across countries and regions. In this paper, we ask whether the application of machine learning models to the prediction of migration intentions can help us understand more about the extent to which migration aspirations across the world are predictable, the different relevance of individual-level migration drivers across countries and regions, and, finally, the comparative performance in predictability of a global predictive model, built on combined data from the whole world, versus the local predictive models, built on specific country data.

One further reason why understanding migration drivers is of high relevance is that policy makers are increasingly trying to tackle the "root causes" of migration (Hagen-Zanker et al. 2023). Understanding which of the 'root causes' is more salient in predicting migration intentions could help in the design of more effective policies (Hagen-Zanker et al. 2023), or, at least, in the identification of relevant factors that are outside policy control. Moreover, there is growing interest in using predictive models to manage migration flows, particularly the inflow of asylum or refugees (e.g., Carammia, Iacus, and Wilkin 2022). Artificial intelligence (AI) and other machine learning technologies that extract sensitive information on a vulnerable population are implicated in some migration management systems (Nalbandian 2022). Concerns about fairness and bias issues of such technologies when delegating for example, asylum decisions or visa allocation on automated or partially automated systems (Molnar 2023; Bircan and Korkmaz 2021), have been raised. Examinations of the performance of these systems are crucial, but we lack access to the data and algorithms used in these systems. Moreover, AI in migration management is likely to deepen global inequalities, especially as least-developed countries lack the infrastructure to implement such systems (McAuliffe 2023). Some migration literature has questioned whether such systems should be built at all (Nalbandian 2022), and for whose benefit are they being built. Examining the feasibility and fairness of such systems is, therefore, crucial, but hindered by the proprietary nature of the data or software, and lack of access to large scale administrative datasets due to privacy concerns.

However, little attention has been paid to the question of what else can we learn about migration as a sociological phenomena by employing such data and methods, while at the same time generating knowledge on the practicality of such techno-social systems in the real world. Accounting for predictive factors of migration aspirations across countries might not only allow us to peek inside these black boxes, and potentially improve these predictive models, it might also highlight important limitations of such approaches when built on similarly large, historical datasets, as well as migration theory. In this sense, machine learning approaches could further our understanding of individual-level factors that inform decision-making in migration, their heterogeneity across countries, and the relevance of country context (Timmerman et al. 2014). However, it is a rather open debate the extent to which human behavior can be predicted on the basis of traditional data sources. For example, in a large scale collaboration, the predictability of six life outcomes among a cohort of US children was found to be rather

poor (Salganik et al. 2020), leading some scholars to speak about general limitations of predictability for human behavior (I. Lundberg et al. 2024). This, however, contrast with the predictability of mortality, where the age at which a person would die was predicted with high accuracy on the basis of detailed, large register data in combination with generative AI in the form of large language models (Savcisens et al. 2024).

Most current machine learning applications to migration have shown limited success in forecasting specific types of mobility, such as asylum and forced displacements (Angenendt, Koch, and Tjaden 2023; Boss et al. 2023; Carammia, Iacus, and Wilkin 2022; Napierała et al. 2022). Other researchers have used Google search queries to predict emigration aspirations (Böhme, Gröger, and Stöhr 2020), tapping into a digital trace left when prospective migrants search for information about potential destinations. However, aggregate models based on aggregate data do not consider how certain changing individual-level factors which act as drivers of migration intentions could impact aggregate flows. In contrast, Molina et al. (2023) used a machine learning principled variable selection process to assess the link between climate change and migration. Previous research has found that migration aspirations are hard to predict in the context of country samples, and that the known migration drivers do not perform well in those study contexts (Ruhnke and Rischke 2024). Other variables were found to be more important than those often touted in the literature, such as social cohesion and political representation (Ruhnke and Rischke 2024).

Migration aspirations have been conceptualized in diverse ways within the literature, reflecting a range of sometimes conflicting theoretical traditions (Carling and Collins 2020; Detlefsen, Heidland, and Schneiderheinze 2022). The connection between migration intentions and behavior is not straightforward, even though these seem to correlate strongly in empirical assessments (Tjaden, Auer, and Laczko 2019). One important finding in this literature is that the correlation seems stronger in some regions and weaker in others, especially so in African countries (Tjaden, Auer, and Laczko 2019; European Commission, Joint Research Centre 2018). According to the migration aspirations and capabilities approach, the ability to migrate, in terms of resources, is as important a determinant as the intentions or aspirations to migrate (Carling and Schewel 2018). National or regional contexts are likely to affect the relevance of individual-level factors in predicting emigration intentions. For example, the social norms and expectations about migration behavior, as well as the opportunities that exist, all depend on the local, regional, national, and international context in which prospective migrants reside (Carling and Schewel 2018). Within a region like Latin America, there are important variations in the levels of migration intentions and its time trends (Sellers 2021). Within specific countries, for example in Senegal, there are susbtantial regional variations in the reasons Senegalese youth claims for wanting to emigrate (Carling et al. 2013). Local labor market variables – for example, the employment and unemployment probabilities, considered as one of the most important factors for economic migration – affect the migration aspirations of youth in the MENA region (Ramos 2019); with worsening employment prospects associated with higher aspirations to emigrate (Dennison 2022). The role of internet access and its use, which likely are function of the country's level of development and infrastructure, have also been shown to affect emigration intentions (Grubanov-Boskovic et al. 2021). Finally, changing political cir-

cumstances in a given country, such as Libya and its civil war, further affect migrants onward migration intentions (Syed Zwick 2022).

Our paper seeks to broaden the scope of migration research globally (Lucassen, Lucassen, and Manning 2010), by exploring how migration intentions differ across various countries in the world and the extent to which these intentions can be predicted by a large set of migration drivers usually studied in the literature. Machine learning can be a helpful tool to enhance our understanding of individual migration decisions, their diversity, and the role of country context (Timmerman et al. 2014). We employ data from the Gallup World Poll (GWP) for the years 2007-2016. The GWP is a world-wide survey covering a large range of topics, in rather small samples in each country, each year. By pooling various survey years we increase the sample size and the number of countries available for our analysis, overcoming the small sample size limitations of previous research. A predictive model based on the algorithm XGBoost is estimated to predict migration intentions (our target variable) on the basis of a large set of factors captured in the survey, and which previous studies have indicated to be predictive of migration intentions or migration behavior (Aslany et al. 2021). XGBoost is a popular algorithm based on a combination of additive trees that learn from the errors in an iterative fashion (Chen and Guestrin 2016). This algorithm is useful for tabular data that performs well in a variety of prediction tasks (Shwartz-Ziv and Armon 2022). After obtaining our predictions, we compute SHAP values (SHapley Additive exPlanations, Sundararajan and Najmi 2020; S. Lundberg 2017) for each predictive feature used in the models. SHAP values are an approach to explain or describe why an algorithm, in this case XGBoost, is making the predictions it makes - an indirect measure of the relevance or salience of a predictive factor. Once we obtain these values for each feature and country, we reduce the dimensionality of this matrix by employing PCA and UMAP to find which countries have similar 'profiles' in terms of their migration aspiration drivers. We describe how these patterns relate to regional variability and to the different levels of migration intentions, and we finalize the paper by highlighting how this approach could be expanded in future research.

Results

Migrant aspiration vary greatly across the world, with some countries having very high levels of migration intentions, and others having very low levels. This is shown in Figure 1, where we map the average migration intentions across the world. The Andean region, Central America and the Caribbean in the Americas, Western and Central Africa, a few salient countries in Western Asia, and Eastern Europe are the regions with the highest levels of migration aspirations. In contrast, South East Asia and high income countries have relatively low levels.

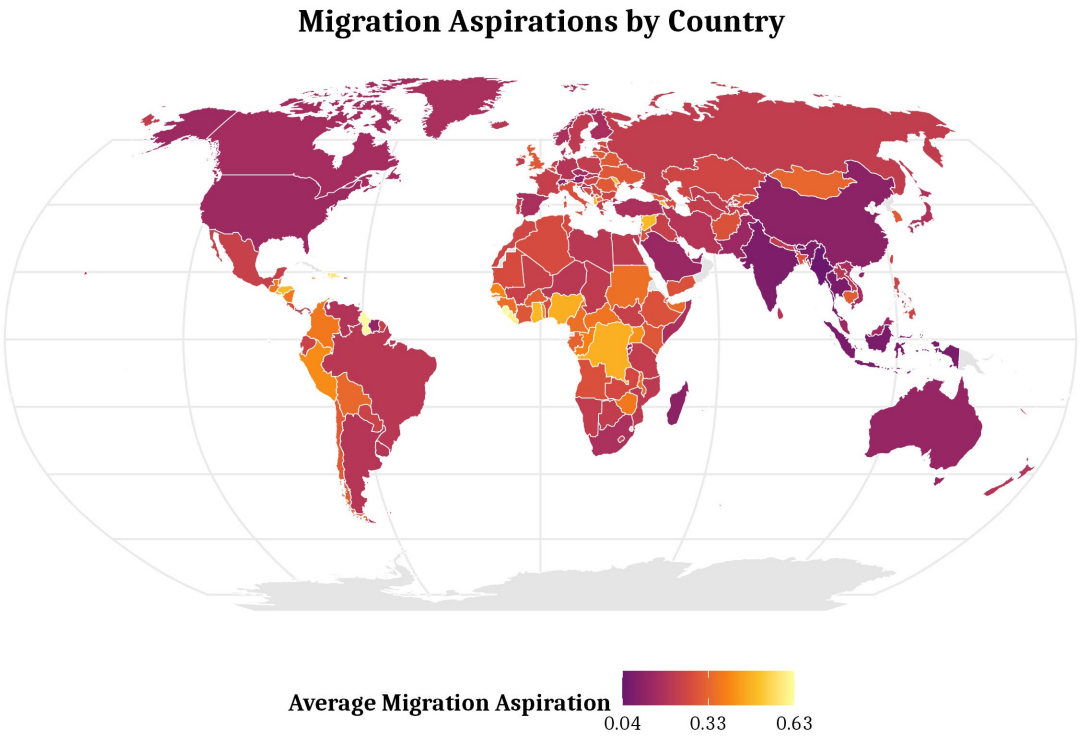

Figure 1: Average migration aspirations by country 2007-2016, GWP

Performance metrics

Variation in migration aspirations across the world could translate into different levels of predictability. To test this empirically, we estimate a global predictive model that pools all countries and years, and a local predictive model for each country - pooling only country specific waves for each country.

Figure 2 shows the accuracy of our global predictive model and the local predictive models for the countries in the GWP, measured in terms of the area under the ROC curve (i.e., AUC). For most countries, the AUC of the global model (i.e., blue dotted line) is higher than the accuracy of the local models (grey bars) with a few exceptions. We see no substantial relation between the accuracy of the local models and the prevalence of the migration intentions in the country (see red circles). Therefore, these results do not simply reflect the prevalence of migration intentions. Both the local and the global models improve on the predictability of migration intentions when compared to the non-informative classifier (or with the prevalence), shown as solid red line in Figure 2.

There are, however, more nuances when comparing the global and local predictive models. In Figure 3, we show the difference between the AUC of the global and local predictive models for each country. Countries where the local model performs better than the global model are colored in red, and countries where the global model performs better are colored in blue. Countries where the global model performs better are mostly high-income countries, such as Canada, Australia, Northern Europe, or even South-East Asia. Countries where local models outperform the global model are mostly concentrated in South Asia and Western, Central, Eastern, and Southern Africa. This comparison is illustrative of the type of biases that a hypothetical global predictive model for managing migration using automated systems could display. Although the global model is built on more than 1.3 million observations, it is not able to capture the heterogeneity in migration drivers in specific countries as well as a local model trained exclusively on data from that country. In so far as these predictive models represent various theories of migration intention formation, we find that such models do not have the same level of predictive power across all countries. In other words, a single theoretical model of migration intentions is likely inadequate to capture the complexity of migration intentions across the world. This is further corroborated when we look at variation in the relevance of migration drivers across countries.

Feature importance

We use SHAP values to assess the importance or relevance of each of the features in predicting migration intentions.

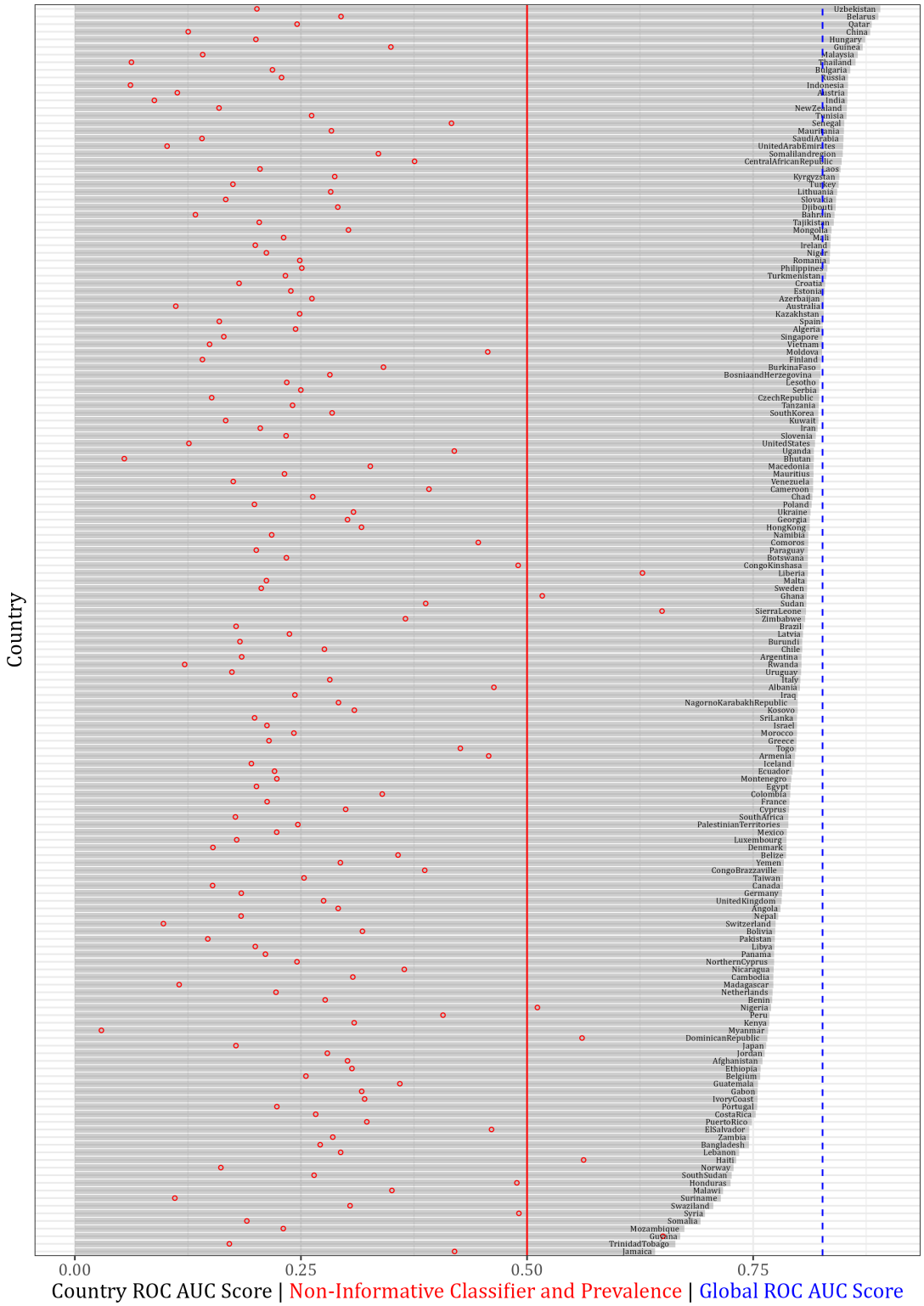

Figure 2: Accuracy of Global and Local predictive models

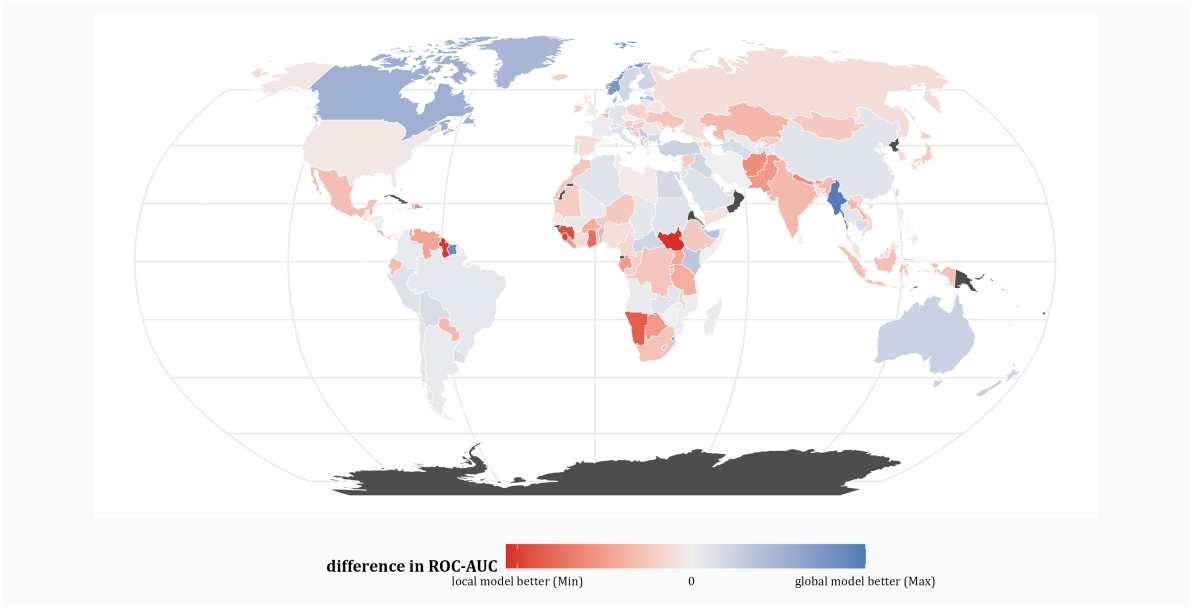

Figure 3: Difference between global and local predictive models' accuracy for each country

Global predictive model

Figure 4 shows the SHAP values for the global predictive model and the top fifty features. We select the best 10 models according to cross-validation to provide a distribution of SHAP values instead of a single value, and plot these employing a boxplot, avoiding in this way the influence of random variation in the feature selection steps of XGBoost. The features age, having relatives or friends abroad, respondents' marital status, and their satisfaction with life in their city build the group of top predictors, with SHAP values substantially higher than those of other predictors. In particular, age and having relatives or friends abroad are by far the main predictors the global predictive model is making use of to generate the predictions. Other predictors have much less salience, meaning their impact on making predictions is much smaller. We also find various income measures among the top fifty predictors, with important variables related to employment or self-employment, as well as respondents' evaluation of major destination countries, such as the US or Germany, as world politics and economic leaders, as well as their perception of the economic conditions in their own countries.

Local predictive models

Figure 5 shows the SHAP values for the local predictive models and, again, the top fifty features. We present this information in terms of a heatmap. Brighter colors, towards yellow, are indicative of higher SHAP values, and hence, higher importance in predicting migration intentions in that country, whereas colder colors, towards dark purple, are indicative of lower importance. We find a substantial similarity in the importance of age, having friends or relatives abroad,

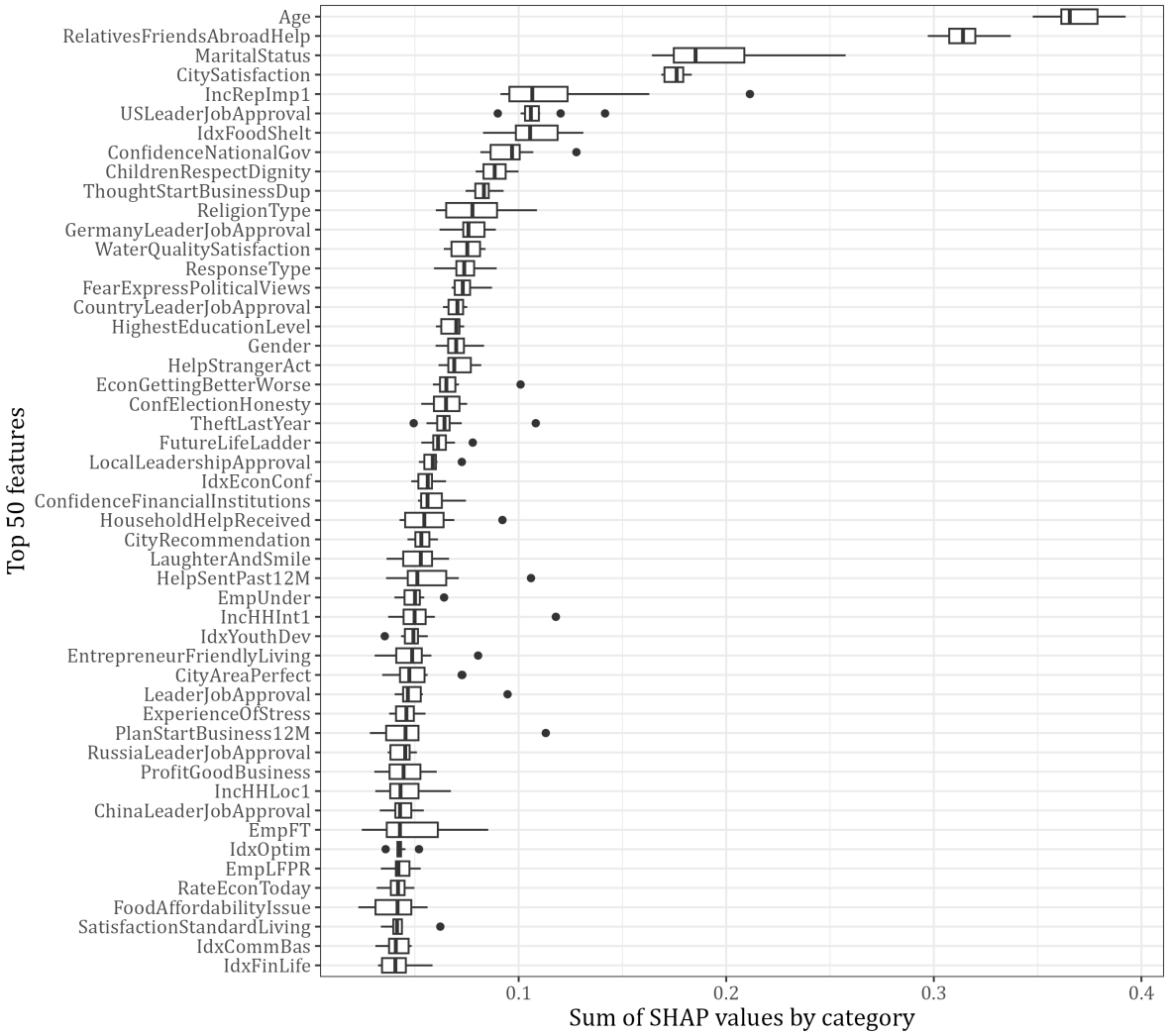

Figure 4: SHAP values for global model

marital status, and satisfaction with life in one's city across almost all countries. We believe this lends credence to the importance of these factors in predicting migration intentions, as captured by previous research (e.g., Aslany et al. 2021) However, the most important lesson of this figure is the substantial heterogeneity in the importance of other migration drivers across countries. For example, gender appears important in some countries (e.g., Niger, Yemen, or Tunisia), but of little relative importance in most countries. Other factors, such as internet connectivity, do not appear predictive at all (Grubanov-Boskovic et al. 2021). This variability is also observed for other features, such as employment, income, or poverty. Importantly, this variability is not well represented in our theoretical understanding of migration intention formation, and could only be discovered by employing a data-driven approach such as the one we present here.

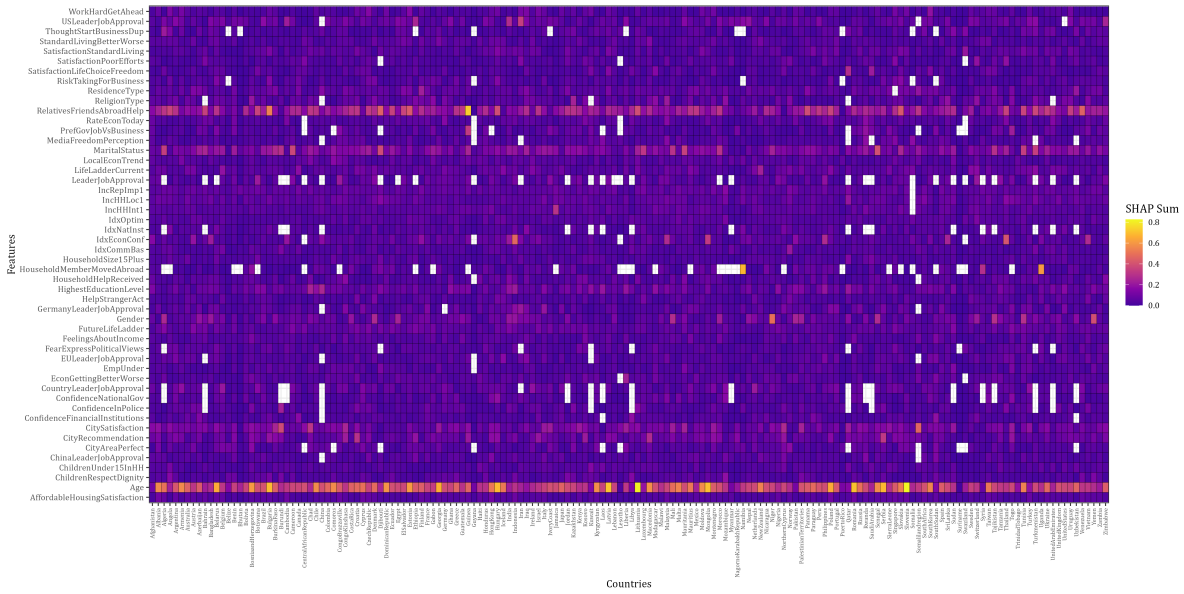

Figure 5: SHAP values across the fifty most important dimensions by country

Dimensionality Reduction and Clustering

To explore whether these patterns in the differential relevance of migration intention predictors relate to regional context, we use a dimensionality reduction technique known as UMAP. Results of UMAP are shown in Figure 6, where each country is represented by two dimensions summarizing the fifty most important SHAP values, after employing PCA on the raw SHAP matrix. The color of each dot corresponds to a world region. We see a substantial degree of clustering by geographic region (e.g., most dots of one color are clustered together), suggesting that the relevance of specific features is associated with regional characteristics. In other words, closer countries in a geographical sense have a similar profile in the predictors of migration as captured by the matrix of SHAP values.

Furthermore, we also see important income level groupings. The upper-left quadrant of Figure 6 is mostly composed of high-income countries from northern and western Europe, Australia and New Zeland, as well as Canada, and the United States, followed by a cluster of high- to mid-income countries in East Asia, Eastern Europe, and the former Soviet Union (SU). In the lower half of Figure 6, in purple, we have cluster of Latin American and Caribbean countries in purple, whereas in green and red, we observe countries from the MENA region and Sub-Saharan Africa. In the upper-left quadrant, in turn, are further African countries (in red), next to a cluster of South and East Asian countries (in gray), and a cluster of mixed countries from multiple regions. These clusters are not perfect. Italy, for example, does not appear near countries with its level of economic development, but we observe a general pattern of clustering by region, which is indicative of the importance of regional context in the formation of migration intentions.

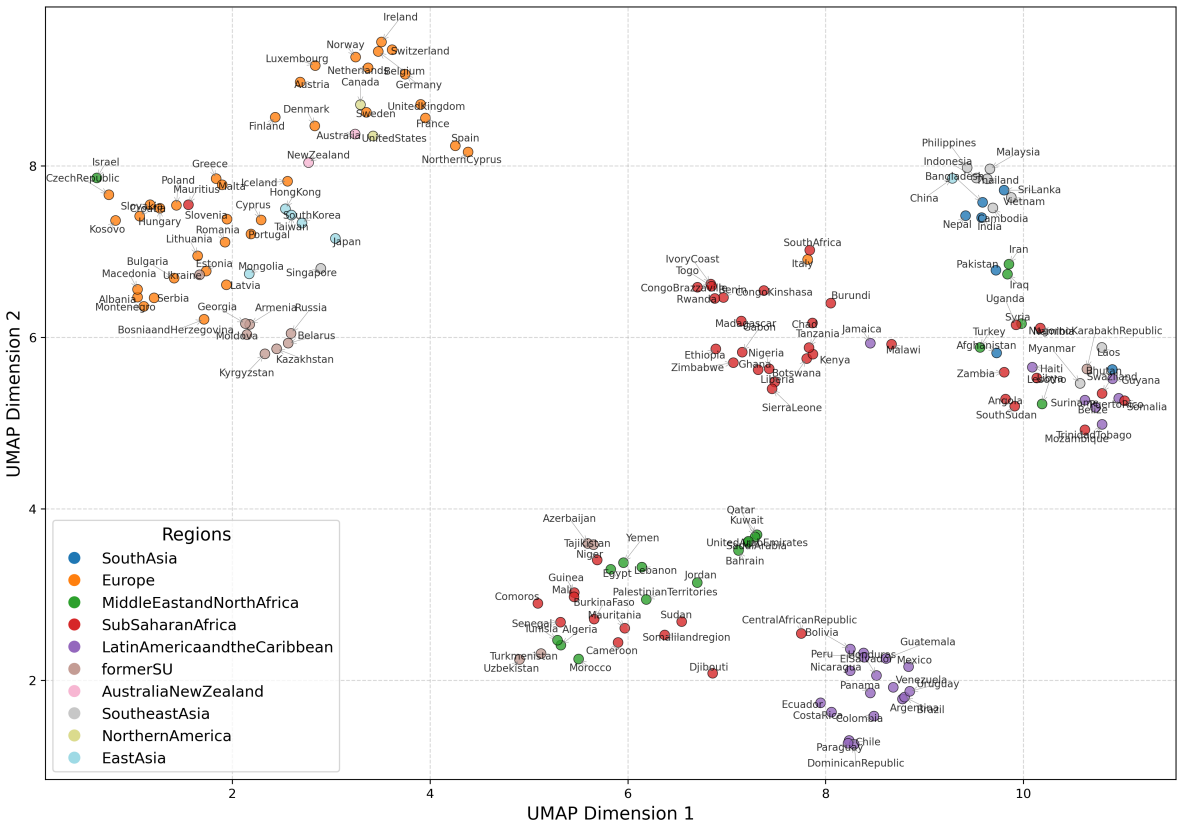

Figure 6: Embedding of fifty most important dimensions according to UMAP

According to these results, there is a clear clustering of countries according to whether they are mostly receiving or sending migrants. Countries hosting large numbers of migrants from all over the world, mainly high-income European and North American ones, appear clearly demarcated from the rest of the world. Mostly migrant-sending countries from the rest of the

world are separated from that cluster, and further form more or less homogeneous specific regional clusters. Both of these results suggest the presence of further structural and region-specific characteristics potentially affecting the formation of migration intentions.

Discussion

In this paper, we have shown that migration aspirations are predictable, to some extent. We reach higher accuracy in predicting migration aspirations, as measured by the AUC, than in other studies (e.g., Molina et al. 2023; Ruhnke and Rischke 2024). We are the first to empirically demonstrate that although a global model can capture country heterogeneity, country-specific models can perform better in lower-income countries, in general, with some important exceptions. This is telling of the inadequacy of a global model, and perhaps also of global theoretical models as well, to capture the complexity of migration intentions across the world, which we show depend substantially on context. For example, we find that the importance of migration drivers varies across countries. And we show that countries in close geographical proximity tend to resemble each other in terms of the importance of migration drivers (i.e., they have similar profiles).

Our results point to a high degree of heterogeneity in the importance of migration drivers across countries that is not well-represented in the theory. This is an important shortcoming that could be addressed by future research. Country context is a salient moderator of how a machine learning predictive model of migration intentions and, by extension, migration behavior or mobility, would operate in real life to forecast or to manage migration flows. Machine learning technologies in migration management should account for country heterogeneity for fairness and bias issues, as well as to potentially improve on predictability. Future research could focus on what drives the similarities/differences between world regions in terms of migration drivers.

Importantly, if country or regional context matter in predicting migration intentions, then algorithmic approaches employed in the management of migration should reflect this. For example, an algorithm that predicts whether someone will emigrate should likely differentiate between countries of origin. If this algorithm is built on data from a single country, for example, or from an average of countries, and is applied in a country that did not contribute to the training data, it is likely to perform poorly in this country. Contrarily, an algorithm built on a rich data set that captures most countries could actually learn more nuances about the relation between migration intentions and individual-level factors, overcoming the limitations of algorithms trained on national data only.

Our approach improves our understanding of future global human mobility patterns (Kraemer et al. 2020), by singling out the most important factors in each context that our model is using. To the extent that our model is capturing important theoretical constructs, this approach constitutes an empirical test of the adequacy of migration drivers. We found a couple of factors that were relevant across almost all context, such as age, having connections or relatives

abroad, being satisfied with life in respondents' city of residence, the marital status, and various measures of income. Most other factors were less relevant, and show greater variability across countries. For example, among these other factors (e.g., gender, employment, or poverty) appear completely irrelevant for prediction. Other types of "attitudes" captured in GWP do not inform the prediction much, though we find perceptions of economic conditions in the country of origin and in large destination countries to be prominent among the most important features (Manchin 2023). Our findings are in line with previous attempts at measuring the relevance of migration drivers employing GWP data (Milasi 2020), but our approach is far more comprehensive and detailed, and we are able to show the relevance of country context in the predictability of migration intentions. In fact, the geographical clustering of countries based on the importance of migration drivers' profiles suggest that the formation of migration intentions might be context specific. One corollary of this is that the extent that migration intentions capture some of the migration behavior (Tjaden, Auer, and Laczko 2019), our findings suggest that actual migration drivers are heterogeneous by country as well. Therefore, information campaigns to inform about the risk of journeys might have to adjust the message to the specifics of a country (Tjaden and Gninafon 2021; Pagogna and Sakdapolrak 2021).

Following the 'complex drivers' conceptualization (Belot and Ederveen 2012), where cultural distance could be considered as the outcome of multiple drivers, our results suggest that interactions between country context and different migration drivers affect our model's predictions. This requires further research but is certain that country context moderate the salience of a specific factor in predicting migration intentions. In fact, world regions, as they are differently embedded in globalized flows of goods, services, and migration systems, also likely shape these patterns. An advantage of our approach in comparison to previous evaluations of migration drivers (e.g., Soto Nishimura and Czaika 2024) lies in the use of a non-parametric approach and model-agnostic evaluation of feature importance (i.e., SHAP values), delinking the assessment of migration driver importance from assumptions about model fit. Classifying migration drivers in a small number of dimensions is useful for theoretical purposes, but tends to obscure the multiple-way interactions between the individual indicators. Our approach, instead, due to its non-parametric nature, flexibly accommodates these interactions, and uses the model-agnostic SHAP metric to unravel the relevance any given feature has in the prediction of migration intentions.

Overall, our machine learning and data-driven perspective shows the relevance of well known factors. For example, demographic factors, such as age and marital status (but also, importantly, potentially period or cohort effects); cumulative models of migration through networks or relatives abroad (Massey et al. 1993); and the emphasis on differentials in socioeconomic conditions (e.g., Van Hear, Bakewell, and Long 2020). Hence, our approach lends support to some level of generalization in migration theory (Carling and Schewel 2018), while at the same time highlighting major shortcomings. In some countries, migration intentions are predicted by other features. A machine learning approach is useful in "sorting" out or ordering various empirical and theoretical findings that need to be put back in the context of theory development. Our findings lead us to ask whether we are observing the truly relevant features of the problem, and what highly predictive features are not being capture by surveys close to the

GWP. Something that is missing in most literature is the relevance of respondents' evaluation of the conditions of migrants in various countries. This is likely an important component that could be easily incorporated in surveys, and which broadly relate to the "pull" effect that conditions in the imagined countries of destination can exert on prospective migrants. More importantly, there is lack of clarity whether the plethora of migration aspiration drivers should be considered as causal mechanisms or simply predictive factors. At the heart of this debate lies the question of how migration aspirations are formed (Carling and Collins 2020; Van Hear, Bakewell, and Long 2020; Beine, Bertoli, and Fernández-Huertas Moraga 2016), which theoretical model underpins current thinking, and, ultimately, what works when trying to predict migration intentions. Whatever factors motivate people to migrate likely depend on important contextual and structural elements that have received less attention in the literature.

Overall, although we make important progress, predicting of such a complex social phenomena like the formation of migration intentions remains a challenging problem. Complex dynamical systems are hard to predict due to the presence of, among others, non-linear effects in social networks (Albert and Barabási 2002). Our ability to predict individual-level intentions, who desires to emigrate, will in general be different than the ability to predict actual migration behavior. It is also a different task than forecasting how many people will migrate (Carammia, Iacus, and Wilkin 2022). Although ours is not a causal model, it is a straightforward test of some previously made theoretical claims. Some of the features that we found to be non-predictive of emigration intentions could, in principle, have a rather small causal effect and remain qualitatively meaningful, but is challenging to provide causal effects for factors with such small associations to emigration intentions. However, our results do relate to the causal inference literature in that with our approach we can highlight the salience of migration drivers that affect selection into migration (Detlefsen, Heidland, and Schneiderheinze 2022), and which could be used to inform future causal models of migration intentions.

With improvements in prediction accuracy, machine learning models could be use to monitor country or regions where data on emigration intentions or behavior is missing. One critical failure of current migration prediction systems is the assumption that migration across countries obeys the same "laws" or theoretical models employed in migration research. Building technological systems for a more equitable migration management requires the inclusion of flexible, data adaptive and contextual models. However, critically examining AI technologies deployed in migration management depends upon the availability of data and the openness of these systems to auditing. This is no easy task when data cannot be publicly accessed due to privacy concerns, and algorithms are proprietary. While we may not have access to the same datasets feeding these systems, we can still use available information to establish empirical boundaries and benchmarks.

Limitations

Our approach does not capture all investigated migration drivers – e.g., Soto Nishimura and Czaika (2024) suggest that more than one-thousand distinct variables have been investigated in the migration determinants literature. This is due to the limited number of variables in the GWP. Increasing the number of drivers could potentially improve the predictive performance

of our models and also change the similarity in country profiles as found in the UMAP analysis, but we argue that our approach is already a substantial improvement over previous studies and hypothesize that the general patterns found in this paper would hold. Our approach is so far the first to attempt to find similarities in migration driver profiles across such a large sample of countries. Other machine learning models could also be used to obtain our predictions (e.g., random forests, as done in Ruhnke and Rischke 2024). We assessed a selection of other models and found XGBoost to be the most accurate, in line with previous studies comparing the performance of this algorithm (Shwartz-Ziv and Armon 2022). The measurement of migration intentions in the GWP is based on three questions, which could be considered a limitation. In fact, our survey items make mention of ideal circumstances in which respondents would like to move to another country, which has considerable limitations (Mjelva and Carling 2023). In this regard, collecting more accurate or theoretically relevant measures of migration intentions could further our understanding of predictability and migration driver importance, but we are limited by the data collected in the GWP. In contrast to previous research where intentions are far better measured (e.g., Ruhnke and Rischke 2024), our data has a higher sample size and a far larger geographical coverage than most previous studies (i.e., a handful of country cases). Although we can capture the role of often not considered individual-level migration drivers, other well known social or aggregate scale drivers (Migali et al. 2018), such as social or economic inequality (Nikolova 2023), at the country level, cannot be examined in this framework. In part, the salience of context suggest that these structural, aggregate or socieltal-level drivers are important for predicting migration intentions accurately, which our findings support.

Future research

So far, the literature of migration aspirations has not considered the role of the intended destination. In fact, theories of migration intentions and migration behavior are not destination specific, and focus far more on motives or reasons to migrate (De Haas, Castles, and Miller 2019). It is possible that our finding of diverse drivers of migration reflect the diversity of destinations. Future work could assess whether aspiring to migrate to different places is driven by different set of drivers. Equally important would be to assess whether these findings can be replicated with better measures of migration intentions or considering the strength of those intentions as well. This work could incorporate features related to those destinations, falling theoretically within the purview of so called "pull factors," something we are unable to do in this paper. Looking at how the predictability of migration intentions evolves is a critical future research question given that the predictability not only differs by region but also over time. Lack of predictability could further signal substantial changes in the formation of migration intentions, for example, during or previous to economic crises. Future research should also check whether aggregate predictions of higher migration intentions match emigration rates from some of these countries in future survey waves, but more data would be needed to perform such analysis.

Materials and Methods

Data

We employed Gallup World Poll (GWP) data for the years 2007-2016 (i.e., amounting to ten cross-sectional waves). The GWP data comes from probabilistic surveys carried out in each country amid the population older than 18 years old. These surveys constitute an important complement to data offered by national statistical offices (Rzepa et al. 2024). Among the questions and items of the survey, we selected almost all the *core* variables that appear in most countries and years. These variables are equally coded across surveys. Our analyses are based on two samples, a Global-GWP, with more than 1.3 million observations, which is the pool of all years and all countries, and a Country-GWP, which pools all years for each country separately - most countries with around 10000 pooled observations. Both the Global-GWP and the Country-specific-GWP are larger samples than those used in previous studies predicting migration aspirations/intentions (e.g., Ruhnke and Rischke 2024).

Migration aspirations/intentions

Our target variable, mobility intentions, was created by combining three different variables from the Gallup World Poll: intentions to study, work, or generally move abroad. For each of these questions, we replaced non-responses like "Don't know", "(Refused)", or empty strings with missing values, and recoded the answers ("Yes" = 1, "No" = 0). We drop all cases for which answers to any of these questions were missing. Finally, we created our binary target variable mobility intentions, assigning the value of one if any of the move to study, work, or move abroad generally were positive, otherwise migration intentions is zero.

Predictors of migration intentions

Features represent the large diversity of migration drivers in the literature (Soto Nishimura and Czaika 2024). These features are summarized in Table 1 in *Supplementary Materials* but include indicators of the most relevant dimensions of migration drivers that have been explored in the literature, taking advantage of the large number of survey items contained in the GWP. For example, thematically, the variables selected include indicators from the major nine dimensions in which migration drivers can be classified, according to Soto Nishimura and Czaika (2024), namely demographic, economic, environmental, human development, individual and household resources, politico-institutional, security, socio-cultural, and supranational. This amounts to approximately 275 variables or raw features - from the core questionnaires, which adds up to 808 features for the model after one-hot encoding of categorical variables.

Workflow

The workflow of our analysis is depicted in Figure 7, where we show the different steps that were taken from the raw data files to the different analyses. This workflow was followed for both

the global predictive model, all pooled observations, and the local predictive models, which are the country sub-samples of the full pooled sample. We start with the raw data files from the GWP. Here we selected some first variables from the core questionnaire, common to all survey years, and then we performed some data cleaning and pre-processing (e.g., making sure all items had the same number of categories). We split the data into a train and test set (i.e., 80/20 split). To each of these samples, we apply various pre-processing steps. We hot-encode all categorical features, splitting a categorical feature with k features into k additional columns. Furthermore, we normalize all numeric variables, and apply a zero variance filter for columns that do not vary across respondents in the respective samples. Here it is worth noting that we excluded any regional or country identifiers from the analyses to avoid capturing patterns related to surveys.

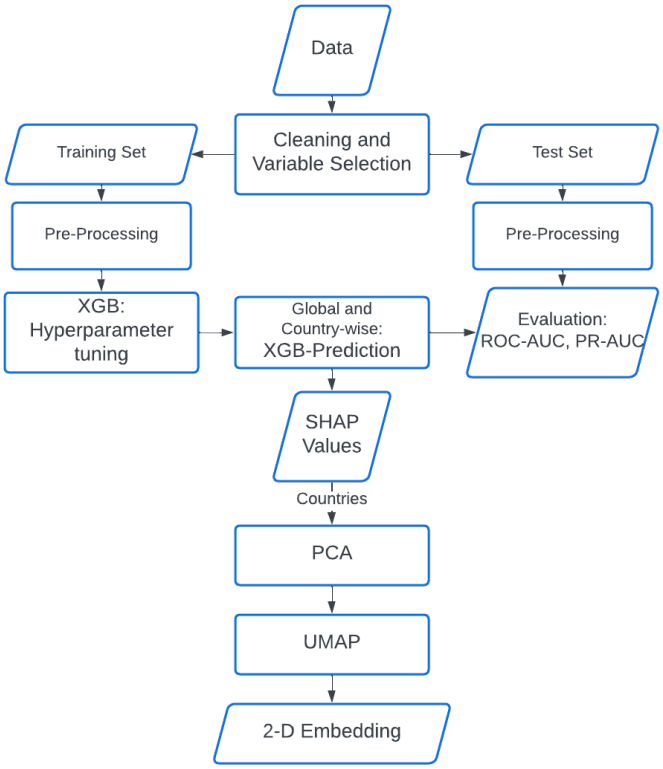

Figure 7: Proposed workflow from data to embedding

XGBoost

To the train data set, we apply the XGBoost algorithm (Chen and Guestrin 2016), a non-parametric model suited for high-dimensional data sets, employing hyperparameter tuning

by means of 5-fold cross-validation. The main idea behind XGBoost is to combine, in an iterative fashion, a series of regression trees or simple learners, that are sequentially fit to the residuals of previous simple learners. The objective function of this algorithm is a combination of a training loss, suited to a binary variable, with a regularization to account for model complexity (i.e., the number of simple learners, regression trees, used in the model) We use the training data (with multiple features) x_i (i.e., our migration drivers) to predict a target variable y_i (i.e., migration aspirations). The hyperparameters we tuned in these models are , and we employ a Bayesian selection method for the search space. In various comparative analysis, XGBoost appears as the leading algorithm in terms of its predictive performance, especially so for tabular data (Shwartz-Ziv and Armon 2022). Other machine learning models could be used for the same exercise, but we choose this algorithm for its performance and ease of use. We employ a bagging approach, combining the best ten models after tuning hyperparameters, using the 10 best-parameter combinations and ensembles these models to create feature importance, SHAP values, and to compute the performance evaluation metrics (e.g., AUC) Evaluation of performance, employ ROC-AUC and PR-AUC, were computed using our models, but predicting only in the test sample with respect to the test sample's true labels.

SHAP values

Once we settle on final best performing models in the train data set, and using the global and local predictive models, we employ SHapley Additive exPlanations (SHAP, Sundararajan and Najmi 2020; S. Lundberg 2017) to these models. Features that were not selected by XGBoost were imputed with the value zero. SHAP values for categorical variables were added up to the original feature to obtain a single value for each feature and compare them to the numeric features. The main idea behind SHAP values is that each feature is assigned an importance value for a particular prediction made. It is game theoretic concept where the features present in a given model are akin to "game players" and the model are the "rules of the game." Changing the features used by a model will impact the model's output, and therefore the accuracy of the predictions; whichever feature is most important in changing the model's output is, therefore, a more important feature. We compute SHAP values for the global predictive model and for the local predictive models, as shown in Figures 4 and 5.

PCA and UMAP

Finally, based on an aggregate data set of all local predictive models' SHAP values, a country by feature dataset, we perform PCA to obtain the main principal components of the SHAP matrix, and then employ UMAP to reduce the dimensionality of this data, as well as to visualize the similarity of countries in terms of SHAP values. PCA is dimensionality reduction method that finds the directions of maximum variance in the data, and projects the data into

these directions. We take the matrix of SHAP values (i.e., all numeric values) and perform PCA on this rectangular matrix, keeping the first 50 dimensions

UMAP, short for Uniform Manifold Approximation and Projection (McInnes, Healy, and Melville 2018), is a non-linear dimensionality reduction technique that constructs a graph in the high-dimensional space of the data and then represents it in a lower dimensional space. The main assumptions behind this methodology are that there is a lower-dimensional representation of the data, and that the distance between data points is meaningful, which we believe are warranted by the application to migration intentions. We test for different number of neighbors and kept the level 5 because of our relatively small data size (i.e., number of countries in the GWP). However, no substantial changes were found in the resulting embedding, see *Supplementary Materials* for more details.

Acknolwedgements

ARS and AM designed the study. ARS wrote first coding of a part of the machine learning analyses, and AM further developed the code and performed further analyses. AM and ARS produced all figures. ARS wrote the first draft of the paper. JT provided access to the Gallup World Data. AM, JT, and ARS revised the paper. All authors approved the final version of the paper.

References

Albert, Réka, and Albert-László Barabási. 2002. "Statistical Mechanics of Complex Networks." *Reviews of Modern Physics* 74 (1): 47.

Angenendt, Steffen, Anne Koch, and Jasper Tjaden. 2023. "Predicting Irregular Migration: High Hopes, Meagre Results."

Aslany, Maryam, Jørgen Carling, Mathilde Bålsrud Mjelva, and Tone Sommerfelt. 2021. "Systematic Review of Determinants of Migration Aspirations." *Changes* 1 (18): 3911–27.

Beine, Michel, Simone Bertoli, and Jesús Fernández-Huertas Moraga. 2016. "A Practitioners' Guide to Gravity Models of International Migration." *The World Economy* 39 (4): 496–512.

Belot, Michèle, and Sjef Ederveen. 2012. "Cultural Barriers in Migration Between OECD Countries." *Journal of Population Economics* 25: 1077–1105.

Bircan, Tuba, and Emre Eren Korkmaz. 2021. "Big Data for Whose Sake? Governing Migration Through Artificial Intelligence." *Humanities and Social Sciences Communications* 8 (1): 1–5.

Böhme, Marcus H, André Gröger, and Tobias Stöhr. 2020. "Searching for a Better Life: Predicting International Migration with Online Search Keywords." *Journal of Development Economics* 142: 102347.

Boss, Konstantin, Andre Groeger, Tobias Heidland, Finja Krueger, and Conghan Zheng. 2023. *Forecasting Bilateral Refugee Flows with High-Dimensional Data and Machine Learning Techniques*. BSE, Barcelona School of Economics.

Carammia, Marcello, Stefano Maria Iacus, and Teddy Wilkin. 2022. "Forecasting Asylum-Related Migration Flows with Machine Learning and Data at Scale." *Scientific Reports* 12 (1): 1–16.

Carling, Jørgen, and Francis Collins. 2020. "Introduction: Aspiration, Desire and Drivers of Migration." In *Aspiration, Desire and the Drivers of Migration*, 1–18. Routledge.

Carling, Jørgen, Papa Demba Fall, María Hernández-Carretero, Mame Yassine Sarr, and Jennifer Wu. 2013. "Migration Aspirations in Senegal: Who Wants to Leave and Why Does It Matter." *European Policy Brief, Brussels, Jan.*

Carling, Jørgen, and Kerilyn Schewel. 2018. "Revisiting Aspiration and Ability in International Migration." *Journal of Ethnic and Migration Studies* 44 (6): 945–63.

Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.

Clemens, Michael, Cindy Huang, Jimmy Graham, and Kate Gough. 2018. "Migration: What You Make It – Seven Policy Decisions That Turned Challenges into Opportunities." Center for Global Development. https://www.cgdev.org/publication/migration-what-you-make-it-seven-policy-decisions-turned-challenges-opportunities.

De Haas, Hein, Stephen Castles, and Mark J Miller. 2019. *The Age of Migration: International Population Movements in the Modern World*. Bloomsbury Publishing.

Dennison, James. 2022. "Re-Thinking the Drivers of Regular and Irregular Migration: Evidence from the MENA Region." *Comparative Migration Studies* 10 (1): 21.

Detlefsen, Lena, Tobias Heidland, and Claas Schneiderheinze. 2022. "What Explains People's Migration Aspirations? Experimental Evidence from Sub-Saharan Africa." *Experimental Evidence from Sub-Saharan Africa (October 5, 2022)*.

European Commission, Joint Research Centre. 2018. *Many More to Come? Migration from and Within Africa*. JRC 110703. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/1702.

Grubanov-Boskovic, Sara, Sona Kalantaryan, Silvia Migali, and Marco Scipioni. 2021. "The Impact of the Internet on Migration Aspirations and Intentions." *Migration Studies* 9 (4): 1807–22.

Hagen-Zanker, Jessica, Jørgen Carling, Ignacio Carrasco, Mathias Czaika, Marie Godin, and Marta Bivand Erdal. 2023. "Tackling the Root Causes of Migration." *Changes* 1: 29.

Kraemer, Moritz UG, Adam Sadilek, Qian Zhang, Nahema A Marchal, Gaurav Tuli, Emily L Cohn, Yulin Hswen, et al. 2020. "Mapping Global Variation in Human Mobility." *Nature Human Behaviour* 4 (8): 800–810.

Lucassen, Jan, Leo Lucassen, and Patrick Manning. 2010. "Migration History: Multidisciplinary Approaches." In *Migration History in World History*, 1–35. Brill.

Lundberg, Ian, Rachel Brown-Weinstock, Susan Clampet-Lundquist, Sarah Pachman, Timothy J Nelson, Vicki Yang, Kathryn Edin, and Matthew J Salganik. 2024. "The Origins of Unpredictability in Life Outcome Prediction Tasks." *Proceedings of the National Academy*

of Sciences 121 (24): e2322973121.

Lundberg, Scott. 2017. "A Unified Approach to Interpreting Model Predictions." *arXiv Preprint arXiv:1705.07874*.

Manchin, Miriam. 2023. "Global Evidence on the Relative Importance of Nonfinancial Drivers of International Migration Intentions." *International Migration Review*, 01979183231162627.

Massey, Douglas S, Joaquin Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J Edward Taylor. 1993. "Theories of International Migration: A Review and Appraisal." *Population and Development Review*, 431–66.

McAuliffe, Marie. 2023. "AI in Migration Is Fuelling Global Inequality: How Can We Bridge the Gap?" World Economic Forum. 2023. https://www.weforum.org/stories/2023/01/ai-in-migration-is-fuelling-global-inequality-how-can-we-bridge-gap/.

McInnes, Leland, John Healy, and James Melville. 2018. "Umap: Uniform Manifold Approximation and Projection for Dimension Reduction." *arXiv Preprint arXiv:1802.03426*.

Migali, Silvia, Fabrizio Natale, Guido Tintori, Sona Kalantaryan, Sanja Grubanov-Boskovic, Marco Scipioni, Fabio Farinosi, et al. 2018. *International Migration Drivers*. EUR 29333 EN JRC112622. Luxembourg: Publications Office of the European Union. https://doi.org/10.2760/63833 (online), 10.2760/565135 (print).

Milasi, Santo. 2020. "What Drives Youth's Intention to Migrate Abroad? Evidence from International Survey Data." *IZA Journal of Development and Migration* 11 (1).

Mjelva, Mathilde Bålsrud, and Jørgen Carling. 2023. "Surveys on Migration Aspirations, Plans and Intentions: A Comprehensive Overview." *Open Research Europe* 3.

Molina, Mario D, Nancy Chau, Amanda D Rodewald, and Filiz Garip. 2023. "How to Model the Weather-Migration Link: A Machine-Learning Approach to Variable Selection in the Mexico-US Context." *Journal of Ethnic and Migration Studies* 49 (2): 465–91.

Molnar, Petra. 2023. "Digital Border Technologies, Techno-Racism and Logics of Exclusion." *International Migration* 61 (5): 307–12.

Nalbandian, Lucia. 2022. "An Eye for an 'i:'a Critical Assessment of Artificial Intelligence Tools in Migration and Asylum Management." *Comparative Migration Studies* 10 (1): 32.

Napierała, Joanna, Jason Hilton, Jonathan J Forster, Marcello Carammia, and Jakub Bijak. 2022. "Toward an Early Warning System for Monitoring Asylum-Related Migration Flows in Europe." *International Migration Review* 56 (1): 33–62.

Nikolova, Milena. 2023. "The Relationship Between Inequality and Potential Emigration: Evidence from the Gallup World Poll." *International Migration Review*, 01979183231202991.

Niva, Venla, Alexander Horton, Vili Virkki, Matias Heino, Maria Kosonen, Marko Kallio, Pekka Kinnunen, et al. 2023. "World's Human Migration Patterns in 2000–2019 Unveiled by High-Resolution Data." *Nature Human Behaviour* 7 (11): 2023–37.

Pagogna, Raffaella, and Patrick Sakdapolrak. 2021. "Disciplining Migration Aspirations Through Migration-Information Campaigns: A Systematic Review of the Literature." *Geography Compass* 15 (7): e12585.

Ramos, Raul. 2019. "Migration Aspirations Among Youth in the Middle East and North Africa Region." *Journal of Geographical Systems* 21 (4): 487–507.

Ruhnke, Simon, and Ramona Rischke. 2024. "Predicting Mobility Aspirations in Lebanon and

Turkey: A Data-Driven Exploration Using Machine Learning." *Data & Policy* 6: e47.
Rzepa, Andrew, Steve Crabtree, Benedict Vigers, and Kiki Papachristoforou. 2024. "Bridging the Gap: Gallup's Role Supporting the Official Statistics Ecosystem." *Statistical Journal of the IAOS* 40 (3): 727–40.
Salganik, Matthew J, Ian Lundberg, Alexander T Kindel, Caitlin E Ahearn, Khaled Al-Ghoneim, Abdullah Almaatouq, Drew M Altschul, et al. 2020. "Measuring the Predictability of Life Outcomes with a Scientific Mass Collaboration." *Proceedings of the National Academy of Sciences* 117 (15): 8398–8403.
Savcisens, Germans, Tina Eliassi-Rad, Lars Kai Hansen, Laust Hvas Mortensen, Lau Lilleholt, Anna Rogers, Ingo Zettler, and Sune Lehmann. 2024. "Using Sequences of Life-Events to Predict Human Lives." *Nature Computational Science* 4 (1): 43–56.
Sellers, Laura Marie. 2021. "Emigration Intentions, Participation Patterns, and Expatriate Voting in Latin America: A Study of People, Politics, and Migration." PhD thesis, Vanderbilt University.
Shwartz-Ziv, Ravid, and Amitai Armon. 2022. "Tabular Data: Deep Learning Is Not All You Need." *Information Fusion* 81: 84–90.
Soto Nishimura, Akira, and Mathias Czaika. 2024. "Exploring Migration Determinants: A Meta-Analysis of Migration Drivers and Estimates." *Journal of International Migration and Integration* 25 (2): 621–43.
Sundararajan, Mukund, and Amir Najmi. 2020. "The Many Shapley Values for Model Explanation." In *International Conference on Machine Learning*, 9269–78. PMLR.
Syed Zwick, Hélène. 2022. "Onward Migration Aspirations and Destination Preferences of Refugees and Migrants in Libya: The Role of Persecution and Protection Incidents." *Journal of Ethnic and Migration Studies*, 1–20.
Timmerman, Christiane, Helene Marie-Lou De Clerck, Kenneth Hemmerechts, and Roos Willems. 2014. "Imagining Europe from the Outside: The Role of Perceptions of Human Rights in Europe in Migration Aspirations in Turkey, Morocco, Senegal and Ukraine." In *Communicating Europe in Times of Crisis: External Perceptions of the European Union*, 220–47. Springer.
Tjaden, Jasper, Daniel Auer, and Frank Laczko. 2019. "Linking Migration Intentions with Flows: Evidence and Potential Use." *International Migration* 57 (1): 36–57.
Tjaden, Jasper, and Horace Gninafon. 2021. "Raising Awareness about the Risk of Irregular Migration: Quasi-Experimental Evidence from Guinea." *Population and Development Review*.
Van Hear, Nicholas, Oliver Bakewell, and Katy Long. 2020. "Push-Pull Plus: Reconsidering the Drivers of Migration." In *Aspiration, Desire and the Drivers of Migration*, 19–36. Routledge.

[Supplementary Materials]- Contextualizing the formation of individual-level migration intentions across the world using machine learning

Alejandra Rodríguez Sánchez Arne Maaß Jasper Tjaden

2024-12-05

This document contains the supplementary materials for the paper "Contextualizing the formation of individual-Level migration intentions across the world using machine learning". We present further descriptive statistics of the variables used in our global and local predictive models, as well as further results on the PCA and UMAP analyses.

Distribution of country sample sizes for the local predictive models

Most of the pooled country samples have a size of 10000 respondents, with few falling below this number of being far higher than this.

Descriptives of the global sample before imputation

Table 1 contains basic descriptive statistics of the different variables

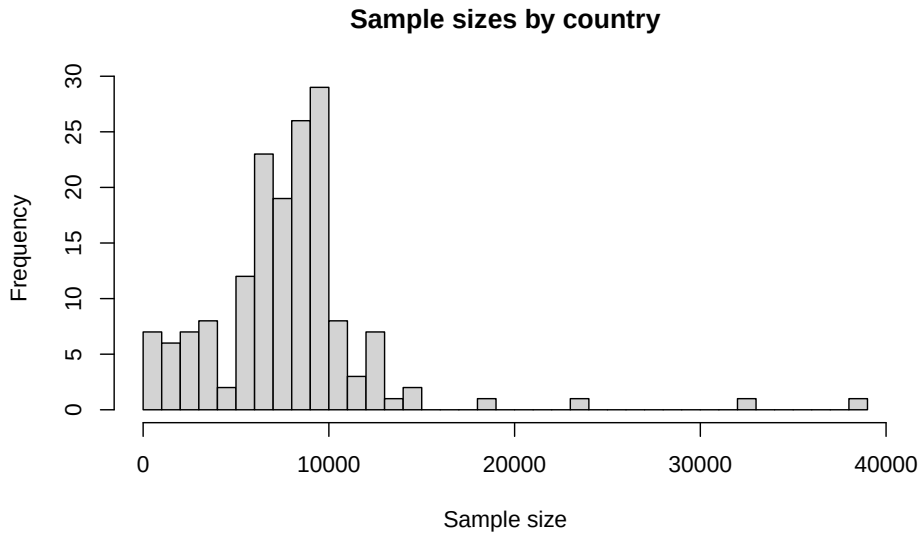

Figure 1: Distribution of sample sizes in the local predictive models

Variable	Descriptives
EmpCat	Employedfulltimeforanemployer:288178, Employedfulltimeforself:153876, Employedparttimedonotwantfulltime:81809, Employedparttimewantfulltime:77849, Outofworkforce:437286, Unemployed:71097, NA:208003
EmpFT	0:384906, IndexScore:288463, NA:644729
EmpFTPop	Fulltimeforanemployer:292606, Notfulltimeforanemployer:836967, NA:188525
EmpLFPR	0:437826, IndexScore:673369, NA:206903
EmpUnder	NotUnderemployed:524013, Underemployed:148989, NA:645096
EmpUnemp	NotUnemployed:602216, Unemployed:71153, NA:644729
HHSize	Mean: 4.23 Median: 4.00 SD: 2.72 Min: 1.00 Max: 22.00 NAs: 1177275
IncHHLoc	Mean: 3361733.01 Median: 72000.00 SD: 16952112.19 Min: 0.00 Max: 1200000000.00 NAs: 1177275
IncPCLoc	Mean: 1061907.62 Median: 22000.00 SD: 5906436.15 Min: 0.00 Max: 360000000.00 NAs: 1177275
IncQuint	Fourth20:29558, Middle20:27092, Poorest20:23540, Richest20:35586, Second20:25047, NA:1177275
IncRepImp	Exactreportedvalue:100503, Imputedfromcategoricalresponse:21191, Imputedwithoutcategoricalresponse:19129, NA:1177275
IdxCommBas	Mean: 59.47 Median: 57.14 SD: 28.93 Min: 0.00 Max: 100.00 NAs: 100130
IdxCivEng	Mean: 32.63 Median: 33.33 SD: 31.76 Min: 0.00 Max: 100.00 NAs: 70323
IdxComm	Mean: 61.70 Median: 50.00 SD: 33.97 Min: 0.00 Max: 100.00 NAs: 110365
IdxCommAcc	Mean: 68.17 Median: 50.00 SD: 35.43 Min: 0.00 Max: 100.00 NAs: 1146844
IdxCommUse	Mean: 90.28 Median: 100.00 SD: 57.01 Min: 0.00 Max: 150.00 NAs: 1146844
IdxCorr	Mean: 68.09 Median: 100.00 SD: 41.66 Min: 0.00 Max: 100.00 NAs: 160840
IdxDayExp	Mean: 70.33 Median: 80.00 SD: 24.42 Min: 0.00 Max: 100.00 NAs: 42217
IdxDivers	Mean: 49.40 Median: 50.00 SD: 35.29 Min: 0.00 Max: 100.00 NAs: 210689
IdxEconConf	Mean: -4.53 Median: 0.00 SD: 70.51 Min: -100.00 Max: 100.00 NAs: 424167
IdxFinLife	Mean: 32.74 Median: 33.33 SD: 29.86 Min: 0.00 Max: 100.00 NAs: 98991
IdxFoodShelt	Mean: 73.42 Median: 100.00 SD: 37.63 Min: 0.00 Max: 100.00 NAs: 28986
IdxJobClim	Mean: 35.10 Median: 0.00 SD: 39.24 Min: 0.00 Max: 100.00 NAs: 116770
IdxLifeEval	Struggling:739630, Suffering:149840, Thriving:311983, NA:116645
IdxLawOrder	Mean: 69.27 Median: 66.67 SD: 31.03 Min: 0.00 Max: 100.00 NAs: 133822
IdxNatInst	Mean: 51.79 Median: 50.00 SD: 35.99 Min: 0.00 Max: 100.00 NAs: 250820
IdxNegExp	Mean: 27.17 Median: 20.00 SD: 29.29 Min: 0.00 Max: 100.00 NAs: 43225
IdxOptim	Mean: 49.04 Median: 50.00 SD: 32.22 Min: 0.00 Max: 100.00 NAs: 67482
IdxPersHealth	Mean: 70.78 Median: 80.00 SD: 28.20 Min: 0.00 Max: 100.00 NAs: 23770
IdxPosExp	Mean: 68.65 Median: 80.00 SD: 28.82 Min: 0.00 Max: 100.00 NAs: 24954

(continued)

Variable	Descriptives
IdxSocLife	Mean: 76.82 Median: 100.00 SD: 31.43 Min: 0.00 Max: 100.00 NAs: 373515
IdxStruggle	Mean: 61.56 Median: 100.00 SD: 48.65 Min: 0.00 Max: 100.00 NAs: 116645
IdxSuffer	Mean: 12.47 Median: 0.00 SD: 33.04 Min: 0.00 Max: 100.00 NAs: 116645
IdxThrive	Mean: 25.97 Median: 0.00 SD: 43.85 Min: 0.00 Max: 100.00 NAs: 117452
IdxYouthDev	Mean: 64.93 Median: 66.67 SD: 35.05 Min: 0.00 Max: 100.00 NAs: 80809
RateEconToday	Excellent:43828, Good:217045, Onlyfair:382269, Poor:275242, NA:399714
WorkSituation	No:675117, Yes:325447, NA:317534
SelfEmployed	No:495532, Yes:241893, NA:580673
JobSatisfaction	Dissatisfied:57295, Satisfied:192651, NA:1068152
CompGeneralView	Hiringnewpeopleandexpandingthesizeofitsworkforce:31110, Lettingpeoplegoandreducingthesizeofitsworkforce:13821, Notchangingthesizeofitsworkforceor:47296, NA:1225871
IdealJob	NoitisNOTideal:106277, Yesitisideal:210345, NA:1001476
DailyEnjoyment	No:20911, Yes:99396, NA:1197791
CleanWaterAccess	No:17654, Yes:102861, NA:1197583
HighEnergyFeel	No:42690, Yes:76768, NA:1198640
CloseContactsFreq	Mean: 7.94 Median: 5.00 SD: 15.52 Min: 0.00 Max: 907.00 NAs: 1207138
WorldChangeBelief	Agree:100331, Disagree:289050, NA:928717
ProfitGoodBusiness	Agree:74706, Disagree:37989, NA:1205403
LeadersRepresent	Agree:56268, Disagree:54284, NA:1207546
IdealLivingPlace	Agree:83621, Disagree:35125, NA:1199352
ParksSatisfaction	Dissatisfied:57806, Satisfied:59656, NA:1200636
MeetPeopleOpportunities	Dissatisfied:209979, Satisfied:699100, NA:409019
JobOpportunitiesAvail	Dissatisfied:298039, Satisfied:153580, NA:866479
LivingPlaceForDisabled	Goodplace:408546, Notagoodplace:310661, NA:598891
MediaFreedomPerception	No:249105, Yes:582483, NA:486510
EnvProblemsExperience	No:77144, Yes:43110, NA:1197844
MoveDueToEnvIssues	No:98774, Yes:13106, NA:1206218
WaterForCrops	Nonotenoughwater:27989, ToomuchwaterDonotread:2789, Yesenoughwater:61340, NA:1225980
WaterForLivestock	Nonotenoughwater:24713, ToomuchwaterDonotread:2276, Yesenoughwater:63052, NA:1228057
PlanToStartBusiness	No:58672, Yes:14837, NA:1244589
CalmnessDaily	No:31548, Yes:88009, NA:1198541
MinorityFriendlyLiving	Goodplace:680857, Notagoodplace:314568, NA:322673
ReligiousMinorityFriendlyLiving	Goodplace:51922, Notagoodplace:18201, NA:1247975
LGBTQFriendlyLiving	Goodplace:306808, Notagoodplace:565987, NA:445303
ImmigrantFriendlyLiving	Goodplace:689347, Notagoodplace:383900, NA:244851
EntrepreneurFriendlyLiving	Goodplace:268489, Notagoodplace:113142, NA:936467
CharityDonationAct	No:863137, Yes:376892, NA:78069

(continued)

Variable	Descriptives
VolunteerAct	No:986861, Yes:252608, NA:78629
HelpStrangerAct	No:637933, Yes:593777, NA:86388
VoiceOpinionAct	No:655560, Yes:169654, NA:492884
ConfidenceInPolice	No:371386, Yes:687690, NA:259022
SafetyAtNight	No:453097, Yes:720162, NA:144839
OptimismPersistence	Agree:158612, Disagree:49043, NA:1110443
GoalPursuitPersistence	Agree:258978, Disagree:72653, NA:986467
NewBizFriendlyCity	Goodplace:526050, Notagoodplace:251090, NA:540958
BizRoleModelPerception	No:111685, Yes:336078, NA:870335
ProfitAsSuccessMetric	Agree:221460, Disagree:75008, NA:1021630
RiskTakingForBusiness	Agree:293060, Disagree:186828, NA:838210
GovEaseOfBizStart	Easy:169800, Hard:247102, NA:901196
GovEaseOfBizManage	No:81786, Yes:52375, NA:1183937
KnowBizAdvicePerson	No:162129, Yes:143646, NA:1012323
OwnBusinessStatus	No:729188, Yes:133075, NA:455835
BusinessEmployeeCount	Mean: 8.68 Median: 6.00 SD: 16.36 Min: 2.00 Max: 907.00 NAs: 1215359
BizMoreMoneyReason	No:6102, Yes:15144, NA:1296852
BizJobLossFearReason	No:18025, Yes:3054, NA:1297019
BizBeBossReason	No:5470, Yes:15827, NA:1296801
BizGreatIdeaReason	No:22480, Yes:48608, NA:1247010
BizFormallyRegistered	No:64750, Yes:56414, NA:1196934
BizTrainingAccess	No:183092, Yes:73573, NA:1061433
BizFundsAccess	No:211357, Yes:56943, NA:1049798
TheftLastYear	No:995987, Yes:181193, NA:140918
AssaultLastYear	No:633445, Yes:49600, NA:635053
ReligionDailyImportance	No:315596, Yes:864748, NA:137754
HouseholdSize15Plus	Mean: 3.14 Median: 3.00 SD: 1.87 Min: 0.00 Max: 97.00 NAs: 40286
ReligiousServiceAttendance	No:205892, Yes:191379, NA:920827
Gender	Female:706241, Male:611856, NA:1
Age	Mean: 40.65 Median: 38.00 SD: 17.33 Min: 13.00 Max: 101.00 NAs: 5123
MaritalStatus	Divorced:43627, Domesticpartner:59431, Married:707009, Separated:28152, SingleNeverbeenmarried:373953, Widowed:91246, NA:14680

(continued)

Variable	Descriptives
WorkCategory	BusinessOwnerStorefactoryplumbingcontractoretcSelfemployed:67218, ClericalorOfficeWorkerInbusinessgovernmentagencyorothertype-ofa0organization97sucha0asatypistsecret:45439, ConstructionorMiningWork-era0Constructionmanagerplumbercarpenterelectricianotherconstructiontradesmi:21619, FarmingFishingorForestryWork-era0Farmerfarmworkeraquacultureorhatcheryworkerfishermandeckhandon:80540, InstallationorRepairWorkerGaragemechaniclinesmanotherinstal-lationmaintenanceorrepairworker:9993, ManagerExecutiveorOfficialInabusinessgovernmentagency-orotherorganization:19816, ManufacturingorProductionWorkerOperatesamachineinafactory-isanassemblylineworkerinafactoryinclu:25871, OtherDonotlist:10596, ProfessionalWorkerLawyerdoctorscientistteacherengineer-nurseaccountantcomputerprogrammer:70397, SalesWorkerClerkinastoredoortodoorsalespersonsalesassociate-manufacturerrepresentativeoutsidesal:35390, ServiceWorkerPolicemanwomanfiremanwomanwaiterorwaitressmaid-nursesaideattendantbarberorbeautic:56779, TransportationWorkerDrivesatrucktaxicabbusetcworkswithoron-aircraftincludingpilotsandflight:15728, NA:858712
GovernmentWorker	No:70206, Yes:22637, NA:1225255
ThoughtStartBusiness	No:87793, Yes:42052, NA:1188253
PlanStartBusiness12M	No:311864, Yes:53324, NA:952910
ChildrenUnder15InHH	Mean: 1.32 Median: 1.00 SD: 1.99 Min: 0.00 Max: 97.00 NAs: 39576
HelpSentPast12M	Both:14555, Livinginanothercountry:20996, Livinginsidethiscountry:111983, Neither:512283, NA:658281
ReligionType	Bahai:189, Buddhist:60999, CaoDai:156, ChineseTraditionalReligionConfucianism:639, Christian:40935, ChristianityEasternOrthodoxOrthodoxyetc:102120, ChristianityProtestantAnglicanEvangelicalSDAsJehovahsWitness-esQuakersAOGMonophysiteAICsPenteco:166470, ChristianityRomanCatholicCatholic:317537, Druze:1235, Hinduism:48281, IslamMuslim:223382, IslamMuslimShiite:19271, IslamMuslimSunni:75857, Jainism:89, Juche:50, Judaism:8565, NeoPaganism:218, Noresponse2011andearlier:11232, Otherlist:7319, PrimalindigeNousAfricanTraditionalandDiasporicAnimistNatureWorshipPaganism:3690, Rastafarianism:277, Scientology:87, SecularNonreligiousAgnosticAtheistNone:77648, Shinto:200, Sikhism:970, Spiritism:1062, TaoismDaoism:2292, Tenrikyo:95, UnitarianUniversalism:148, Zoroastrianism:56, NA:147029
ThoughtStartBusinessDup	No:183747, Yes:96187, NA:1038164

(continued)

Variable	Descriptives
PlanStartBusiness12MDup	No:251781, Yes:28287, NA:1038030
CountryCurrentLadderStep	Mean: 6.18 Median: 6.00 SD: 2.09 Min: 1.00 Max: 11.00 NAs: 1148616
CountryPastLadderStep	Mean: 5.92 Median: 6.00 SD: 2.16 Min: 1.00 Max: 11.00 NAs: 1152532
BizSoleOrPartners	Nohavepartners:8708, Yessoleowner:31455, NA:1277935
BizEmployeeTrend	Decrease:3408, Increase:17949, Staythesame:39549, NA:1257192
CountryFutureLadderStep	Mean: 7.35 Median: 8.00 SD: 2.55 Min: 1.00 Max: 11.00 NAs: 1166427
WorkHardGetAhead	No:233304, Yes:987016, NA:97778
ChildrenRespectDignity	No:383860, Yes:810855, NA:123383
ChildrenLearningOpportunity	No:364199, Yes:861522, NA:92377
SatisfactionPoorEfforts	Dissatisfied:588383, Satisfied:388882, NA:340833
LeaderJobApproval	Approve:299181, Disapprove:225977, NA:792940
SatisfactionEnvEfforts	Dissatisfied:558848, Satisfied:606081, NA:153169
SatisfactionJobIncreaseEfforts	Dissatisfied:279203, Satisfied:152566, NA:886329
SatisfactionLifeChoiceFreedom	Dissatisfied:331361, Satisfied:882108, NA:104629
ConfidenceMilitary	No:272231, Yes:703370, NA:342497
ConfidenceJudicialSystem	No:489486, Yes:512016, NA:316596
ConfidenceNationalGov	No:488354, Yes:505753, NA:323991
ResidenceType	Aa0largecity:420372, Aa0smalltownorvillage:386269, Aa0suburbofalargecity:120508, Aruralareaoronafarm:330485, NA:60464
ConfidenceHealthcare	No:62462, Yes:112884, NA:1142752
ConfidenceFinancialInstitutions	No:423318, Yes:653839, NA:240941
FearExpressPoliticalViews	Mean: 2.53 Median: 2.00 SD: 1.04 Min: 1.00 Max: 4.00 NAs: 581676
DailyFear	No:172969, Yes:27221, NA:1117908
ConfReligiousOrg	No:147593, Yes:436727, NA:733778
EULeaderJobApproval	Approve:388657, Disapprove:266727, NA:662714
ConfMediaIntegrity	No:155815, Yes:189764, NA:972519
ConfElectionHonesty	No:504570, Yes:481177, NA:332351
DailyEnjoyment1	Mean: 3.73 Median: 4.00 SD: 1.17 Min: 1.00 Max: 5.00 NAs: 897881
DailyLearning	Mean: 3.45 Median: 4.00 SD: 1.27 Min: 1.00 Max: 5.00 NAs: 899217
SupportEncourageHealth	Mean: 3.94 Median: 4.00 SD: 1.18 Min: 1.00 Max: 5.00 NAs: 898335
PositiveEnergyFromOthers	Mean: 3.93 Median: 4.00 SD: 1.15 Min: 1.00 Max: 5.00 NAs: 898453
FinancialFreedom	Mean: 2.58 Median: 3.00 SD: 1.33 Min: 1.00 Max: 5.00 NAs: 898341
WorriesAboutMoney	Mean: 3.14 Median: 3.00 SD: 1.49 Min: 1.00 Max: 5.00 NAs: 898777
ActiveProductive	Mean: 3.53 Median: 4.00 SD: 1.23 Min: 1.00 Max: 5.00 NAs: 899069
PhysicalHealth	Mean: 3.64 Median: 4.00 SD: 1.24 Min: 1.00 Max: 5.00 NAs: 897440
CityAreaPerfect	Mean: 3.74 Median: 4.00 SD: 1.25 Min: 1.00 Max: 5.00 NAs: 897928
RecognitionForCityImprovement	Mean: 2.22 Median: 2.00 SD: 1.38 Min: 1.00 Max: 5.00 NAs: 907027
PlanStartBusiness	No:144162, Yes:25443, NA:1148493
TrustForBusinessPartnership	No:81501, Yes:59496, NA:1177101
AccessToBusinessTraining	No:116187, Yes:51283, NA:1150628
AccessToBusinessFunds	No:124748, Yes:43824, NA:1149526
PrefGovJobVsBusiness	Business:200952, Government:187923, Neither:69422, NA:859801

(continued)

Variable	Descriptives
CorruptionInBusiness	No:257796, Yes:811837, NA:248465
CorruptionInGov	No:239182, Yes:791376, NA:287540
EconConditionsGood	Nonotgood:87641, Yesgood:60627, NA:1169830
EconGettingBetterWorse	DonotreadThesame:224479, Gettingbetter:399628, Gettingworse:438975, NA:255016
CountryLeaderJobApproval	Approve:497208, Disapprove:481853, NA:339037
USLeaderJobApproval	Approve:522986, Disapprove:352094, NA:443018
UKLeaderJobApproval	Approve:257725, Disapprove:183189, NA:877184
GermanyLeaderJobApproval	Approve:465408, Disapprove:265335, NA:587355
FranceLeaderJobApproval	Approve:216942, Disapprove:140474, NA:960682
RussiaLeaderJobApproval	Approve:333677, Disapprove:389956, NA:594465
ChinaLeaderJobApproval	Approve:402065, Disapprove:350198, NA:565835
JapanLeaderJobApproval	Approve:188628, Disapprove:96562, NA:1032908
InternetUsage	No:11003, Yes:75735, NA:1231360
CanadaLeaderJobApproval	Approve:116706, Disapprove:54008, NA:1147384
LifeLadderCurrent	Mean: 6.44 Median: 6.00 SD: 2.27 Min: 1.00 Max: 11.00 NAs: 15546
SatisfactionEducationOpportunities	Dissatisfied:37279, Satisfied:78336, NA:1202483
TeachersRespectDignity	No:43293, Yes:73234, NA:1201571
SatisfactionLocalSchools	Dissatisfied:34259, Satisfied:82991, NA:1200848
SatisfactionCountrySchools	Dissatisfied:33442, Satisfied:78613, NA:1206043
InternetAccess	No:83112, Yes:87006, NA:1147980
IndiaLeaderJobApproval	Approve:98489, Disapprove:102432, NA:1117177
LifeLadderPast	Mean: 6.22 Median: 6.00 SD: 2.34 Min: 1.00 Max: 11.00 NAs: 1153545
HomeLandline	No:220087, Yes:127174, NA:970837
PersonalMobilePhone	No:40476, Yes:193412, NA:1084210
FutureLifeLadder	Mean: 6.85 Median: 7.00 SD: 2.25 Min: 1.00 Max: 10.00 NAs: 132498
LifePurposeMeaning	No:8184, Yes:103384, NA:1206530
SatisfactionPersonalHealth	Dissatisfied:151902, Satisfied:549984, NA:616212
HealthProblemsRestriction	No:953043, Yes:311405, NA:53650
FeelingsAboutIncome	Mean: 2.65 Median: 3.00 SD: 0.95 Min: 1.00 Max: 4.00 NAs: 30912
VacationPlansAbroad	No:54293, Yes:8015, NA:1255790
RelativesFriendsHelpAvailability	No:241117, Yes:1000763, NA:76218
SatisfactionCurrentHousing	Dissatisfied:20499, Satisfied:67979, NA:1229620
SatisfactionStandardLiving	Dissatisfied:474919, Satisfied:740748, NA:102431
StandardLivingBetterWorse	DonotreadThesame:364716, Gettingbetter:533197, Gettingworse:321689, NA:98496
HighestEducationLevel	HighestEducationLevel Completedelementaryeducationorlessupto8yearsofbasiceducation:431491, Completedfouryearsofeducationbeyondhighschoolandor-receiveda4yearcollegedegree:196809, Beyondsecondaryeducation915yearsofeducatio:659487, NA:30311
HouseholdCellPhone	No:192844, Yes:859900, NA:265354

(continued)

Variable	Descriptives
RelativesFriendsAbroadHelp	No:707413, Yes:367960, NA:242725
HomeLandlinePhone	No:476963, Yes:417738, NA:423397
HomeElectricity	No:15120, Yes:143376, NA:1159602
HomeTelevision	No:190560, Yes:1018087, NA:109451
HomeComputer	No:108227, Yes:69406, NA:1140465
HomeInternetAccess	No:749039, Yes:458690, NA:110369
FoodAffordabilityIssue	No:896861, Yes:388783, NA:32454
ShelterAffordabilityIssue	No:1002959, Yes:278017, NA:37122
FamilyHungerIssue	No:299358, Yes:41893, NA:976847
HouseholdMemberMovedAbroad	No:238634, Yesreturned:12817, Yesstillthere:38551, NA:1028096
EnvGroupActivity	Nohavenotdone:79061, Yeshavedone:12116, NA:1226921
RecyclingActivity	Nohavenotdone:66383, Yeshavedone:25464, NA:1226251
EnvProductAvoidance	Nohavenotdone:50260, Yeshavedone:40436, NA:1227402
WaterUseReduction	Nohavenotdone:32847, Yeshavedone:59031, NA:1226220
ClimateChangeAwareness	Ihaveneverheardofit:89656, Iknowagreatdealaboutit:51945, Iknowsomethingaboutit:198071, NA:978426
TemperatureRiseOpinion	Aresultofhumanactivities:136216, Aresultofnaturalcauses:50295, BothDONOTREAD:44574, NA:1087013
LocalTempChanges	Colder:5235, Haventlivedherelongenough:233, Stayedaboutthesame:6764, Warmer:33091, NA:1272775
ClimateChangeThreat	Mean: 3.11 Median: 3.00 SD: 0.88 Min: 1.00 Max: 4.00 NAs: 1147890
MilitaryCivTargetJustify	Depends:26307, Neverjustified:193997, Sometimesjustified:53611, NA:1044183
IndivCivTargetJustify	Depends:24026, Neverjustified:218487, Sometimesjustified:46667, NA:1028918
PeacefulMeansEffectiveness	PeacefulmeansALONEwillNOTwork:84567, PeacefulmeansALONEwillwork:151912, NA:1081619
BornInCountry	Borninanothercountry:65894, Borninthiscountry:1182758, NA:69446
JobSatisfactionOld	Dissatisfied:13986, Satisfied:60829, NA:1243283
ClimateChangeImpact	Agree:74322, Disagree:28411, Donotknow:11592, NA:1203773
WaterScarcityConcern	Agree:82371, Disagree:39546, Donotknow:8479, NA:1187702
ExtremeWeatherPerception	Agree:75616, Disagree:30377, Donotknow:8329, NA:1203776
WorkStrengthUse	No:17883, Yes:55731, NA:1244484
WorkEncouragement	No:14135, Yes:23236, NA:1280727
WorkTimeWaste	No:16853, Yes:5506, NA:1295739
OpinionsValuedAtWork	No:4094, Yes:17991, NA:1296013
DesireForRepeatDays	No:55549, Yes:138908, NA:1123641
WellRested	No:409353, Yes:873187, NA:35558
TreatedWithRespect	No:163679, Yes:1120545, NA:33874
ControlOverTime	No:34852, Yes:96086, NA:1187160
LaughterAndSmile	No:357571, Yes:892708, NA:67819
PrideInAchievements	No:29841, Yes:45087, NA:1243170

(continued)

Variable	Descriptives
LearnedSomethingInteresting	No:613360, Yes:671903, NA:32835
GoodTastingFood	No:19181, Yes:56760, NA:1242157
ExperienceOfEnjoyment	No:384930, Yes:907438, NA:25730
ExperienceOfPain	No:910361, Yes:373787, NA:33950
ExperienceOfHappiness	No:209022, Yes:490474, NA:618602
LeaderApprovalCurrent	Approve:221910, Disapprove:145450, NA:950738
ExperienceOfWorry	No:823160, Yes:460644, NA:34294
ExperienceOfSadness	No:1003932, Yes:278759, NA:35407
ExperienceOfStress	No:842490, Yes:372799, NA:102809
ExperienceOfBoredom	No:65062, Yes:22039, NA:1230997
ExperienceOfDepression	No:340553, Yes:53051, NA:924494
IdealNumberOfChildren	Mean: 2.93 Median: 3.00 SD: 1.75 Min: 0.00 Max: 100.00 NAs: 1068164
ExperienceOfAnger	No:1017517, Yes:248632, NA:51949
ResponseType	Rural:349013, Urban:244654, NA:724431
ExperienceOfLove	No:24672, Yes:51215, NA:1242211
SocialTimeWithFamilyFriends	Mean: 5.54 Median: 4.00 SD: 5.01 Min: 0.00 Max: 60.00 NAs: 1174787
TrustForBusinessPartnership1	No:281209, Yes:257274, NA:779615
EaseOfObtainingBusinessLoan	No:191137, Yes:93178, NA:1033783
GovtEaseOfBusinessPaperwork	No:194054, Yes:119902, NA:1004142
ConfidenceInBusinessSuccess	No:64599, Yes:100588, NA:1152911
BizRulesStability	No:65580, Yes:45342, NA:1207176
BizAssetSafety	No:159515, Yes:208592, NA:949991
GovtProfitInterference	No:164850, Yes:187732, NA:965516
FreeTimeHours	Mean: 3.88 Median: 3.00 SD: 3.77 Min: 0.00 Max: 24.00 NAs: 1280771
BizEmployeeConfidence	No:84989, Yes:192806, NA:1040303
PhysicalActivity	No:33046, Yes:18728, NA:1266324
SmokingBehavior	No:72416, Yes:27305, NA:1218377
CitySatisfaction	Dissatisfied:260837, Satisfied:954072, NA:103189
CityImprovementPerception	Gettingbetter:203120, Gettingworse:73979, Thesame:105188, NA:935811
CityRecommendation	Nowouldnotrecommend:321075, Yeswouldrecommend:848706, NA:148317
LocalEconConditions	Nonotgood:275724, Yesgood:294136, NA:748238
LocalEconTrend	DonotreadThesame:304799, Gettingbetter:471327, Gettingworse:388394, NA:153578
PrimaryHeatingSource	CentralheatfromadistanceDistrictheatingenergygrid:7796, Charcoal:3023, Coal:4767, Dung:555, Electricity:12293, Grass:156, NAIdontheatmyhome:31755, Oil:4251, OtherDonotlist:859, Paraffin:735, Peat:73, Propaneornaturalgas:12980, Solar:3293, StrawFirewood:6624, Wind:76, Wood:9682, NA:1219180
LocalLeadershipApproval	Approve:325155, Disapprove:210451, NA:782492
RetirementStatus	No:34322, Yes:15585, NA:1268191
CleanWaterAccess1	No:78658, Yes:35797, NA:1203643

(continued)

Variable	Descriptives
TrustInPeople	Noyouhavetobecarefulindealingwithpeople:133108, Yesmostpeoplecanbetrusted:42158, NA:1142832
DivineInvolvementBelief	IdontbelieveinGod:913, No:39179, Yes:57459, NA:1220547
GovtEmissionReduction	Nonotdoingenough:28385, NotapplicableThisisnotaconcerninthiscountry:1492, Yesdoingenough:12753, NA:1275468
JobSatisfaction1	Dissatisfied:15704, Satisfied:48775, NA:1253619
HealthImpactOnActivities	Mean: 2.58 Median: 0.00 SD: 6.02 Min: 0.00 Max: 41.00 NAs: 1226102
WomenRespectDignity	No:269039, Yes:576238, NA:472821
RecentEmploymentStatus	No:89461, Yes:40374, NA:1188263
SelfEmploymentStatus	No:65386, Yes:29525, NA:1223187
HouseholdHelpReceived	Both:19023, Livinginanothercountry:53988, Livinginsidethiscountry:122418, Neither:846956, NA:275713
PublicTransportationSatisfaction	Dissatisfied:465663, Satisfied:703272, NA:149163
RoadsHighwaysSatisfaction	Dissatisfied:552967, Satisfied:643668, NA:121463
EducationalSystemSatisfaction	Dissatisfied:395705, Satisfied:792800, NA:129593
AirQualitySatisfaction	Dissatisfied:302819, Satisfied:896426, NA:118853
WaterQualitySatisfaction	Dissatisfied:376503, Satisfied:823474, NA:118121
HealthcareAvailabilitySatisfaction	Dissatisfied:508561, Satisfied:672561, NA:136976
AffordableHousingSatisfaction	Dissatisfied:570535, Satisfied:568935, NA:178628
CityBeautySatisfaction	Dissatisfied:244065, Satisfied:637307, NA:436726
IncHHLoc1	Mean: 16622.42 Median: 7659.57 SD: 79086.42 Min: -1.00 Max: 70061250.00 NAs: 407109
IncHHInt1	Mean: 18515.21 Median: 8102.08 SD: 136206.21 Min: 0.00 Max: 83856324.81 NAs: 240591
IncRepImp1	Mean: 6539.16 Median: 2211.73 SD: 44446.08 Min: 0.00 Max: 35030625.00 NAs: 245911
MIG_ASP	Mean: 0.25 Median: 0.00 SD: 0.43 Min: 0.00 Max: 1.00 NAs: 0
YEAR_WAVE	Mean: 2011.74 Median: 2012.00 SD: 2.56 Min: 2007.00 Max: 2016.00 NAs: 0

(continued)

Variable	Descriptives
Country	Afghanistan:10667, Albania:6541, Algeria:6006, Angola:3750, Argentina:9912, Armenia:8881, Australia:9161, Austria:8918, Azerbaijan:8669, Bahrain:9890, Bangladesh:12083, Belarus:8786, Belgium:7938, Belize:891, Benin:5906, Bhutan:2962, Bolivia:8879, BosniaandHerzegovina:6724, Botswana:6959, Brazil:11145, Bulgaria:7653, BurkinaFaso:7858, Burundi:2990, Cambodia:9901, Cameroon:9094, Canada:6028, CentralAfricanRepublic:1995, Chad:7894, Chile:9154, China:32336, Colombia:9941, Comoros:5992, CongoBrazzaville:5854, CongoKinshasa:6639, CostaRica:9881, Croatia:6621, Cyprus:5463, CzechRepublic:6731, Denmark:7669, Djibouti:2984, DominicanRepublic:8929, Ecuador:9006, Egypt:18782, ElSalvador:9839, Estonia:6919, Ethiopia:3947, Finland:6694, France:9662, Gabon:5942, Georgia:8886, Germany:14371, Ghana:8874, Greece:7936, Guatemala:9819, Guinea:5979, Guyana:479, Haiti:3883, Honduras:9765, HongKong:5235, Hungary:7796, Iceland:3053, India:38394, Indonesia:12036, Iran:7965, Iraq:11702, Ireland:8435, Israel:9666, Italy:9891, IvoryCoast:4944, Jamaica:1462, Japan:12971, Jordan:12755, Kazakhstan:9540, Kenya:8941, Kosovo:6524, Kuwait:9888, Kyrgyzstan:9851, Laos:2962, Latvia:6754, Lebanon:12942, Lesotho:998, Liberia:5749, Libya:1985, Lithuania:7279, Luxembourg:6927, Macedonia:6547, Madagascar:6008, Malawi:7977, Malaysia:9002, Mali:6962, Malta:6986, Mauritania:10713, Mauritius:2976, Mexico:9673, Moldova:8641, Mongolia:8786, Montenegro:7726, Morocco:8945, Mozambique:2937, Myanmar:5090, NagornoKarabakhRepublic:995, Namibia:1966, Nepal:10998, Netherlands:7691, NewZealand:6959, Nicaragua:9854, Niger:8910, Nigeria:9671, NorthernCyprus:3944, Norway:3955, Pakistan:14311, PalestinianTerritories:12911, Panama:9820, Paraguay:8895, Peru:8861, Philippines:9947, Poland:8557, Portugal:8904, PuertoRico:472, Qatar:3949, Romania:7818, Russia:23096, Rwanda:5977, SaudiArabia:12806, Senegal:8850, Serbia:7660, SierraLeone:6882, Singapore:10231, Slovakia:6763, Slovenia:7445, Somalia:1875, Somalilandregion:5998, SouthAfrica:8926, SouthKorea:9863, SouthSudan:2824, Spain:9949, SriLanka:10079, Sudan:6308, Suriname:486, Swaziland:994, Sweden:7607, Switzerland:4452, Syria:8429, Taiwan:6874, Tajikistan:9777, Tanzania:8904, Thailand:10988, Togo:3842, TrinidadTobago:1482, Tunisia:10127, Turkey:9835, Turkmenistan:5916, Uganda:7933, Ukraine:8659, UnitedArabEmirates:11684, UnitedKingdom:13344, UnitedStates:7023, Uruguay:8916, Uzbekistan:7929, Venezuela:8848, Vietnam:10804, Yemen:9907, Zambia:6867, Zimbabwe:8879

(continued)

Variable	Descriptives
REG_GLOBAL	AustraliaNewZealand:16120, CommonwealthofIndependentStates:119626, EastAsia:76065, EuropeanUnion:228721, EuropeOther:57126, LatinAmericaandtheCaribbean:180292, MiddleEastandNorthAfrica:190174, NorthernAmerica:13051, SouthAsia:99494, SoutheastAsia:80961, SubSaharanAfrica:256468

Note: Own compilation of variables.

PCA

The first fifty components allow us to explain 80% of the variation in the data, with only the first component explaining around 15%. We consider the first fifty principal components to explain most the variation, as can be seen in the scree plot, Figure 2.

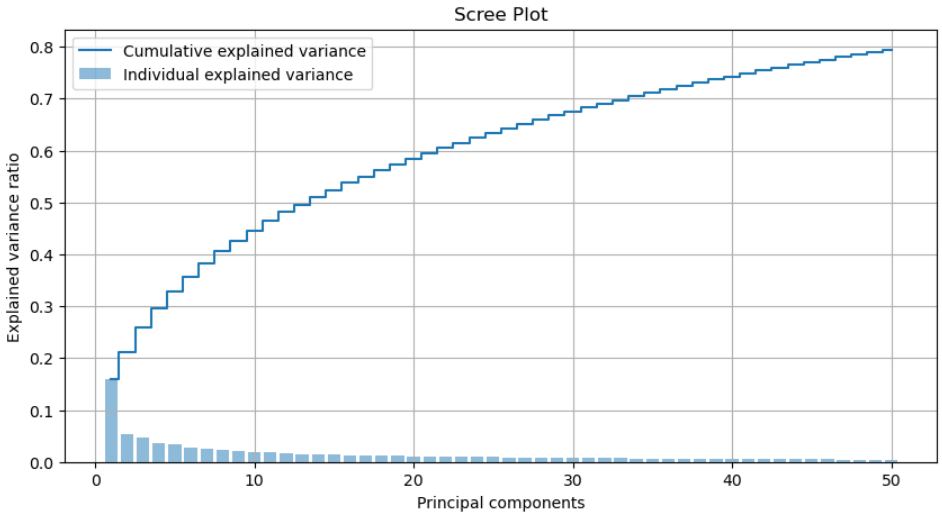

Figure 2: Principal Components and explained variance

UMAP with different neighbors

One of the most important hyperparameters in UMAP is the number of neighbors. Figure 3 shows the UMAP solutions for different numbers of neighbors. We see some qualitative changes but similar patterns across the different solutions, especially in the clustering by regions. Our selected number of neighbors (i.e., five) is hence sufficient to represent the structure in the data in a lower dimension.

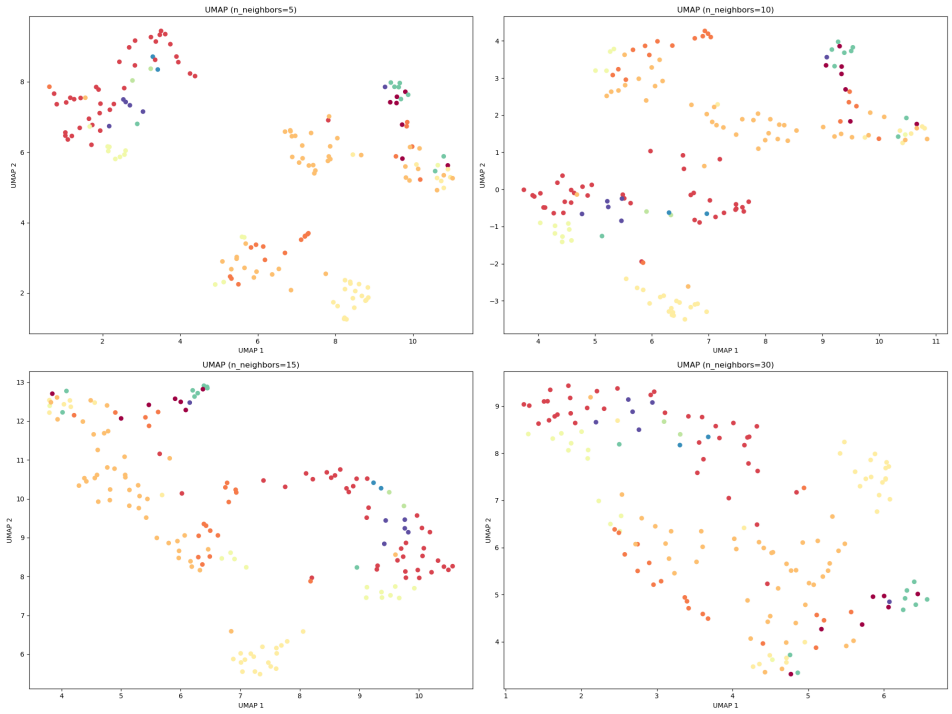

Figure 3: UMAP solutions with different number of neighbors