IFPRI Discussion Paper 01953

July 2020

The Migration Propensity Index

An application to Guatemala

Francisco Ceballos

Manuel A. Hernandez

Markets, Trade, and Institutions Division

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

The International Food Policy Research Institute (IFPRI), a CGIAR Research Center established in 1975, provides research-based policy solutions to sustainably reduce poverty and end hunger and malnutrition. IFPRI's strategic research aims to foster a climate-resilient and sustainable food supply; promote healthy diets and nutrition for all; build inclusive and efficient markets, trade systems, and food industries; transform agricultural and rural economies; and strengthen institutions and governance. Gender is integrated in all the Institute's work. Partnerships, communications, capacity strengthening, and data and knowledge management are essential components to translate IFPRI's research from action to impact. The Institute's regional and country programs play a critical role in responding to demand for food policy research and in delivering holistic support for country-led development. IFPRI collaborates with partners around the world.

AUTHORS

Francisco Ceballos (f.cebillos@cgiar.org) is a Research Fellow in the Markets, Trade, and Institutions Division of the International Food Policy Research Institute (IFPRI), Washington, DC.

Manuel A. Hernandez (m.a.hernandez@cgiar.org) is a Senior Research Fellow in IFPRI's Markets, Trade, and Institutions Division, Washington, DC.

Notices

¹ IFPRI Discussion Papers contain preliminary material and research results and are circulated in order to stimulate discussion and critical comment. They have not been subject to a formal external review via IFPRI's Publications Review Committee. Any opinions stated herein are those of the author(s) and are not necessarily representative of or endorsed by IFPRI.

² The boundaries and names shown and the designations used on the map(s) herein do not imply official endorsement or acceptance by the International Food Policy Research Institute (IFPRI) or its partners and contributors.

³ Copyright remains with the authors. The authors are free to proceed, without further IFPRI permission, to publish this paper, or any revised version of it, in outlets such as journals, books, and other publications.

Abstract

International migration has grown rapidly over the past two decades, at an annual rate of 2.4%, prompting increased interest in identifying the root causes of outmigration and the population groups more likely to emigrate. However, anticipating migration is a complex task, as the decision to migrate is often determined by multiple push and pull factors that are typically interrelated and are not always directly observable. This study proposes the Migration Propensity Index (MPI), a novel approach to indirectly estimate a household's propensity or probability to emigrate. The central idea is to identify and keep track of a reduced set of household-level indicators that are strongly correlated with the (latent) decision of individuals to emigrate. Taken together and converted into an index, the combined indicators reflect the objective likelihood that one or more individuals from a given household will emigrate. The MPI is concise, easy to implement, and statistically rigorous, and avoids asking direct, sensitive questions about migration attempts or intentions, which are prone to refusals and underreporting. We calibrate the index to data for Guatemala, relying on an out-of-sample cross validation procedure using a panel dataset of 2,798 households living in what are considered "vulnerable" municipalities. The data were collected in 2012, 2013, and 2014. We discuss the index design and implementation, including concrete examples of its application. The resulting model includes 12 simple variables (and two location shifters) and correctly identifies 93% of eventual emigrating and non-emigrating households. The MPI can serve policymakers in getting better insights in drivers of migration, monitor present and expected migratory flows, and for targeting of economic and social policies.

Keywords: International migration; Scoring index; Migrants; Migration probability; Guatemala

Acknowledgments

We thank the valuable research assistance of Cynthia Paz. The study was funded by the United States Agency for International Development (USAID). The funding source had no role in data collection, data analysis and report writing. The contents and conclusions of the study are the responsibility of the authors and do not necessarily reflect the views of the International Food Policy Research Institute (IFPRI), USAID or the United States Government.

1. Introduction

International migration has grown rapidly over the past two decades, at an annual rate of 2.4%. According to a recent report from the United Nations (2019), there are about 272 million international migrants across the globe (3.5% of the world population), compared to 174 million in 2000 (2.8% of the population). Western Europe and the United States are the regions that receive most emigrants. 51% of migrants are men and roughly one third have between 15 and 34 years old. In addition, a large share of emigrants comes from rural areas, which receive around 40% of international remittances (IFAD, 2016; FAO, 2018).

Migration is a multidimensional phenomenon determined by a wide set of factors, including push factors, encouraging people to move out of their current location, and pull factors, attracting people to move into a new location (Rubenstein, 2017). These factors are generally grouped into four categories: economic factors (e.g., job opportunities, wages); environmental factors (e.g., food availability, weather); social factors (e.g., availability of services, quality of life); and safety/cultural factors (e.g., political stability, crime). Among these potential factors, The World Bank (2018) emphasizes the role of employment opportunities, wage differentials, and distance (whether physical or cultural) as important global factors shaping the observed international migration patterns. In the case of migration from the Northern Triangle (El Salvador, Guatemala, and Honduras) to the United States, additional factors include family reunification, vulnerability to natural disasters, and crime and insecurity (Cohn et al., 2017; Congressional Research Service, 2019a & 2019b; National Immigration Forum, 2019).¹

Cross-border migration tends to stir political controversy and yield strongly opposing perspectives among main stakeholders, which are the migrants themselves, the families and actors in the communities that they leave behind, and different actors in the destination communities (Usher, 1977). For successful migrants, migration can result in higher incomes, access to services, and improved livelihoods, including better education and nutrition for their children (FAO, 2018); these benefits, however, can take time to materialize, leaving migrants in vulnerable positions (Mattoo, Neagu, and Ozden, 2005). In addition, migration generally occurs at great peril, with many migrants undergoing extreme hardships and even losing their lives in the journey (Donnelly and Hagan, 2014; Isacson, Meyer, and Morales, 2014). In terms of the families that migrants leave behind, remittances can constitute an important source of income allowing them to invest more in education and housing, and attain an overall better quality of life (Yang and Martinez, 2006; Adams and Cuecuecha, 2010). Nevertheless, migration may entail an increased work burden for family members staying behind, such as in agricultural, home-maintenance, and child-rearing activities (Singh, 2015) and a brain drain for the overall communities of origin (Docquier and Marfouk, 2006). Moreover, migration may bring about an array of psychological costs for both migrants and their families (Torres and Wallace, 2013; Torres et al., 2016). The effect on the destination communities is also ambiguous. On one hand, significant dissent can be observed, where immigration is regarded as a threat to local livelihoods, potentially crowding-out local labor supply, disrupting customs, or negatively impacting the safety of communities (Galor and Stark, 1991). On the other hand, others argue that migration can actively contribute to progress and the transformation of societies, both as a current and as a historical phenomenon (De Haan, 1999; FAO, 2018).

All in all, migration reflects both challenges and opportunities and brings about benefits and costs for all actors involved. The objective in this paper is to develop a methodology that allows to monitor the overall

¹ A study from Creative Associates International (2019) identifies economic (unemployment, especially among youth; household economic hardship), transnational ties (having a relative living abroad; receiving remittances), and victimization (crime victim or relative/friend killed) as the three main drivers shaping international migration from the Northern Triangle.

intention of households to permanently migrate to another country and to identify the households more likely to do so in the short term. In doing this, we do not intend to imply that migration is per se a negative phenomenon or one that should be curtailed. Rather, we aim to provide a useful policy tool that can summarize different key factors involved in this decision —without attempting to individually identify or quantify them— and as such contribute to better understand the migration decision, anticipate it, and aid in the devise of policies that can improve the social and economic environment at the country of origin and reduce the incentives to embark on such a risky endeavor.

Identifying migration drivers requires a comprehensive and integrated analysis as migration decisions are generally influenced by many factors at the individual, household, local, regional, and national level. These factors are typically interrelated, vary over time, often reinforce one another, and are not always directly observable, adding complexity to the analysis. For instance, the decision to migrate is not necessarily an individual decision but is likely made together with other family members (Stark, 1991). Migration may also range from voluntary to forced migration (FAO, 2018). Similarly, the potential opportunities of an individual to migrate may depend on his network outside the community and abroad (Munshi, 2003; McKenzie, 2006).² Recent studies further argue that migration may be an adaptation strategy to climate change (Jessee et al., 2018) or a response strategy to natural disaster shocks where migrants' networks play an important role (Mahajan and Yang, 2020), and that higher education and credit constraints may not always drive international youth migration (Valentine et al., 2017; de Brauw, 2019).

We propose and describe the methodological steps to derive and implement a Migration Propensity Index (MPI). The goal of the index is to provide an objective measure of the probability that individuals from a given household will migrate to a foreign country within the next 12 months. As opposed to standard self-reported migration indicators that are potentially subjective, inaccurate, and difficult to monitor over time, the MPI:

1. Avoids sensitive direct questions about attempts and intentions to emigrate, which are prone to refusals or underreporting or may have undesired consequences if they are repeatedly asked over time to the same group of people,³
2. Is easy to implement and can be calculated in the field by relying on a concise set of simple, non-invasive questions that are easy to collect;
3. Is statistically rigorous as it is derived using advanced statistical methods, that account for the timing of migration decisions and the potential factors correlated with these decisions, and tested through cross-validation procedures focusing on the out-of-sample predictive power of the index; and
4. Can aid both monitoring and targeting purposes, including monitoring (direct or indirect) beneficiaries of programs and/or targeting social and economic programs to new households at risk to emigrate.

The proposed MPI is in essence similar to the Progress Out of Poverty/Poverty Probability Index (PPI), originally developed by the Grameen Foundation and currently implemented by Innovations for Poverty Action (IPA) in several countries.⁴ The PPI relies on ten specific household characteristics and asset ownership questions to estimate the likelihood that a household is living below the poverty line (Schreiner,

² See Skeldon (1997) for an extensive review of the determinants of migration.

³ For example, repeated sensitive questions to beneficiaries of a certain program every 3 or 4 months may result in people considering migration more seriously or opting out of the program.

⁴ <https://www.povertyindex.org/>.

2013). Such a tool allows for targeting of social programs by field staff in real time, eliminating the need for lengthy, in-depth income or expenditure surveys or more complex poverty mapping exercises. Naturally, the increased simplicity and immediacy of the approach could come at the cost of reduced accuracy.

Understanding migration decisions, however, imposes different and additional challenges given the nature and potential factors driving these decisions. In particular, the decision to emigrate should be regarded as a latent (hidden) decision process that can be correlated with multiple factors, some of which ultimately push an individual to migrate; and these conditions or circumstances that trigger someone to emigrate may (or may not) periodically change over time. Accordingly, we modify the PPI approach to model the latent emigration decision, with the goal of obtaining a reliable, simple, concise, and statistically sound migration scoring index that can help donors, policy makers, and program implementors better approximate and track the probability of emigration among the population of interest.

We estimate the index in the case of Guatemala, relying on an extensive panel dataset of households in vulnerable (less developed) municipalities collected between 2012 and 2014. Guatemala is one of the three Northern Triangle countries, which exhibit the highest emigration rates in Central America. Recent estimates suggest that around 265,000 people permanently left the Northern Triangle region during each of the last five years, most of them to the United States (Congressional Research Service, 2019b). The latest Population Census in Guatemala indicates that close to 230,000 people emigrated from Guatemala over the last decade (National Statistical Institute-INE, 2018). While we focus on Guatemala, the index can be easily adapted to and implemented in the two other Northern Triangle countries: El Salvador and Honduras. It is important to note that the index should be regarded as a dynamic tool, optimized to the best currently available data but prone to continuous validations and further refinements as more data become available.

The remainder of the note is organized as follows. Section 2 provides details about the data used to calibrate the index. Section 3 discusses the methodology followed to derive the index, including the statistical methods and validation procedures used. Section 4 presents the resulting MPI model and Section 5 describes the steps to implement the index for monitoring and targeting purposes. Section 6 provides concluding remarks.

2. Data

Data Source

To develop the index, we work with a panel dataset of households in Guatemala collected throughout three survey rounds between 2012 and 2014 as part of the monitoring and impact evaluation of the Plan del Pacto Hambre Cero (PPH0).⁵ The survey was designed by the International Food Policy Research Institute (IFPRI) in coordination with the Secretary of Food Security and Nutrition (SESAN), and implemented by the National Statistical Institute (INE). While the PPH0 focused on the 176 municipalities with the highest stunting rates in the country, the panel dataset includes a subsample of 60 municipalities where a specific number of households were followed over time.⁶ These households have children under

⁵ The PPH0 was a national strategy designed by the Government at the beginning of 2012 to address the problem of chronic malnutrition, acute malnutrition, and food insecurity with special emphasis on households with children under five years old, pregnant women and women of childbearing age (between 15 and 49 years old).

⁶ Besides food insecurity and malnutrition, these 176 municipalities also fall behind in a host of other development indicators such as prevalence of poverty, access to basic services, or market accessibility, reason why we refer to them as “vulnerable” municipalities.

five years old, pregnant women, and/or women between 15 and 49 years old, which were the targeted population of the PPH0. As a result, the final panel dataset is representative of this subpopulation within the selected municipalities.⁷ Appendix A includes additional details about the sampling procedure and survey protocols.⁸

The surveys were implemented during the months of November and December of 2012, 2013, and 2014. They collected information on household socioeconomic characteristics, household income sources, expenditures, and assets, dwelling characteristics, access to services and credit, participation in social programs, hygiene and family health, nutrition, and household food consumption. The final panel dataset includes 1,344 households interviewed during both 2012 and 2013 and 1,455 households interviewed during both 2013 and 2014, making a total of 2,798 households (with 779 households interviewed during all three years).

Compared to other available datasets in Guatemala, which are nationally representative but are only cross sectional (i.e. do not follow the same individuals or households over time), the panel nature of the data used in the analysis allows for an increased accuracy to account for the timing of the emigration decisions within a household and the factors correlated with these decisions, thereby avoiding potential confounding effects. Figure 1 provides a hypothetical example that shows the advantages of using the 2012-2014 PPH0 panel compared to the 2014 Encuesta Nacional de Condiciones de Vida (ENCOVI). In the example, the household head emigrates between 2012 and 2013 and begins sending remittances about a year later. Under this scenario, the panel nature of the PPH0 data would permit to capture household characteristics and conditions in 2012, prior to the migration event, allowing a precise identification of the drivers triggering the decision at the time it was made. The cross-sectional nature of the ENCOVI data, instead, would identify the emigrant in the household but would only be able to observe household characteristics after the migrant started sending remittances, confounding the identification of any potential causal effects. The same limitations would apply for the Encuesta Nacional de Salud Materno Infantil (ENSMI) and Sistema de Vigilancia Epidemiológica de Salud y Nutrición (SIVESNU), since these are also not longitudinal. More generally, the use of panel data permits a more refined analysis as it accounts for both inter-household differences and intra-household dynamics as opposed to cross-sectional data, which only accounts for the former.

⁷ In practice, this is very close to the households' universe given that almost 89% of households in the listing had at least one member in these targeted groups.

⁸ For monitoring purposes, the annual surveys also covered other municipalities (communities) that were part of 176 targeted municipalities by the PPH0. These other communities, however, were randomly selected every year and were not followed over time.

Figure 1: Advantage of using panel over cross-sectional data: PPH0 vs ENCOVI

Note: Unfinished floor includes earth, sand or mud floor.

Universe of variables

Considering all the information collected in the PPH0 surveys, the first step was to construct a broad set of variables that could potentially be correlated with the decision to emigrate by an individual in a household. These variables are related to common drivers of emigration, including economic, environmental, social, and safety/cultural factors, generally used in the migration literature. The resulting variable list, together with each variable's definition and type of measure, is presented in Appendix B. These include a total of 48 variables related to household socioeconomic characteristics and composition, dwelling materials, access to services, assets, landholdings and agricultural activities, expenditures, participation in social programs, climate vulnerability, whether someone emigrated during the past year, incoming remittances, and geographic location.

The list is exhaustive of all variables likely associated with the decision to emigrate that are available in the PPH0 surveys, and that should be relatively easy to collect in future surveys (as discussed in the next Section). We recognize, however, that there are still some potentially relevant variables missing in this list, such as those related to whether a household or individual was a victim of crime, employment status of household members, or economic, social, and political perceptions, which were (unfortunately) not collected in the PPH0 surveys. In this line, the model proposed in this report should be viewed as a first attempt to construct a robust MPI. As panel data with a wider set of indicators become available, the

analysis could be further refined to incorporate additional indicators potentially correlated with the decision to emigrate.⁹ Likewise, a panel with a larger sample size (when available), may also help improve the accuracy of the index.

While many of the variables considered are related to local (push) factors, there is a small subset of variables aimed at capturing certain external (pull) factors. Specifically, we consider whether the household knows anyone in their community that emigrated during the previous year or whether other households in the community receive remittances, as well as whether the household itself receives remittances, which turn out to be important predictors of the decision to emigrate. We interpret these variables as capturing the effect of being in contact with an individual from the community who has successfully migrated abroad (i.e. peer or network effects), which is known to be an important pull factor that may prompt other individuals in the community to follow suit.

In terms of broader pull factors, such as immigration policies or job opportunities abroad, we argue that, while important, they generally exert a blanket effect on attracting migrants, affecting all households in a similar manner and thus acting as an overall (positive or negative) shifter to the probability to emigrate. As such, the MPI remains a valuable tool for identifying those households most prone to emigrating within a given group, independent of the current external conditions.

Main migration indicator

As discussed above, the advantage of counting with a panel dataset is that we can observe individual and household characteristics before a household member decides to emigrate. In order to capture whether any individuals migrated in a given household, we rely on the following question asked in each survey round:

- Did any household member leave the community to seek work either temporarily or permanently during the last 12 months?
 - *Yes / No*
- **[If “Yes”]** Where did he/they go?
 - *Same municipality / Same department / Another department / Guatemala City / Another country*

We identify a household with an emigrant during the corresponding year if the answers to the above questions are “Yes” and “*Another country*”. This indicator is used as the dependent variable to calibrate the statistical models and derive their predictive power. Given that the question in the survey refers to the previous 12 months from the date of the interview and that the time between survey rounds was also 12 months, the methodology estimates the probability that a member in a given household emigrates during the 12 months following the date of the interview, controlling for the factors available at the time of the interview.

⁹ Certainly, there can still be unobservable factors specific to each household that could trigger the decision of an individual to migrate, which cannot be accounted for (at least not directly). This limitation, however, is inherent in any analysis of this type regardless of the data used.

3. Methodology

The overall goal of the study is to derive a simple, concise, and statistically-rigorous index based on easily-measurable, reliable household characteristics and conditions that can predict with reasonable confidence whether an individual in a household will emigrate or not in the following 12 months.

Given that the target group to use this index are donors, policy makers, and program implementors and that the index is basically meant for monitoring and targeting as close to real time as possible, the index must be *simple*. In order to be readily adopted, practitioners must be able to grasp the inner workings of the index and trust its capacity to predict a phenomenon they already know reasonably well in qualitative terms. Complex statistical or machine learning models with non-linear effects and cross-interactions between variables are not always transparent or well suited to understanding and require an excessive amount of training data. In this regard, the MPI is based on widely-used statistical regression techniques and focuses on certain key variables that would affect the propensity of an individual to emigrate.

In addition, information gathering in remote, rural areas is costly and faces numerous challenges. Conducting long household surveys in person may be cost-prohibitive in many cases. The index must thus be *concise*. The MPI comprises twelve variables carefully selected to maximize its predictive capacity based on currently available data. Moreover, these variables must be *easy to measure* through direct, non-invasive questions and with minimal reporting bias. For instance, directly asking individuals about their intentions to emigrate may possibly lead to underreporting —out of fear or mistrust—, or to refusals to respond. The MPI must use indicators that can be measured with ease in the field and indirect enough to encourage honest (unbiased) responses. It must also contemplate the possibility of asking these questions over the phone, in order to significantly reduce costly transportation and logistics involved in sending a survey team to the field.

Lastly, if the index is to be used with confidence to guide policy and intervention targeting, it must be *statistically rigorous*. Since the decision to emigrate is dynamic in nature and since this is generally followed by an influx of remittances into the migrant's household, model fit and likely confounding effects (endogeneity problems) must be carefully considered to obtain a rigorous index with a high predictive power. The MPI circumvents endogeneity problems by relying on a large, detailed panel survey of households from vulnerable regions of Guatemala, which, as discussed above, permits to compare a household's *ex ante* characteristics with their *ex post* decision to emigrate. Further, to avoid potential issues around model overfitting, the MPI is constructed and evaluated using cross-validation techniques, which split the working sample into two subsamples: one subsample for model fitting and the other subsample for evaluating the model's predictive capacity.

The next subsections describe the methodology for the construction and validation of the proposed MPI.

The model

Since the dependent variable, whether an individual in the household emigrated during the last 12 months, is binary, we must rely on binary or discrete choice statistical models. We work with a standard logistical regression model (Logit model), which allows to bound the outcome of the model to a probability range between 0 and 1.¹⁰ Since the propensity to emigrate may vary according to local economic, social, and infrastructural conditions we must also control for these aggregated effects (shifters). Given that the

¹⁰ The Logit model is defined as $P(Y = 1|X) = \frac{\exp(X\beta)}{1 + \exp(X\beta)}$ where $P(Y = 1|X)$ is the probability of observing an emigrant in a household ($Y=1$) conditional on a set of X factors. The parameters β are estimated via maximum likelihood estimation. For additional details see Greene (1993).

PPH0 surveys do not cover the entire country and are not representative at the department level, we include regional fixed effects in the specifications. In particular, we include a separate dummy for each of the following four regions or composite regions (for those cases in which the number of observations for a few departments was too small to consider them as a separate region):

- Region 1: Centro + Pacífico
- Region 2: Altiplano Occidental
- Region 3: Corredor Seco + Izabal
- Region 4: Verapaces

All specifications also include an indicator variable to distinguish rural from urban areas, according to the official municipality classification where the household is located. Finally, since some households are included twice in the working sample (i.e. when the household was interviewed across all three survey rounds), we cluster standard errors at the household level to account for the likely correlation in the observed outcomes within the same household.

Out-of-sample cross validation

We rely on an out-of-sample cross-validation procedure to select the final model. This is a standard statistical procedure to assess the predictive performance of a model. The idea is to assess how the model will perform when applied to new data. To achieve this, the data is randomly partitioned into two subsamples: 70% of the observations are used for estimating the model (i.e. model training); and these coefficients are applied to the remaining 30% of the sample to test whether the model predictions match with the actual observations (i.e. model predictive performance). To minimize sampling bias, the procedure of splitting the sample is repeated 300 times.¹¹ In each of these repetitions, the coefficients of the model are estimated with 70% of the observations and the metrics on the model's predictive accuracy are calculated over the remaining 30% of observations. For each metric, we average the measures obtained across the 300 repetitions to derive an overall metric. We consider two metrics for the model's construction:¹²

Root mean square error (RMSE): A usual metric to evaluate the fit of a statistical model, which is constructed by squaring the difference between the model's predicted probability and the observed outcome (either 0 or 1), averaging these across all observations, and taking the square root. It measures how far the model predictions are from the actual occurrences.¹³

Concordance statistic (C-Stat): This metric is commonly used for binary-choice models since it measures the concordance between the observed outcome and the model's predicted probability for that outcome across all observations. A model that maximizes this metric exhibits the most internally-coherent array of predicted probabilities, in the sense that observations with a positive outcome (households having a migrant in our case) will have on average a higher predicted probability than those with a negative outcome.¹⁴

¹¹ These partitions naturally maintain the full-sample proportions of households with and without a migrant during the last 12 months.

¹² We avoid using other standard metrics for the model construction, such as the correct classification rate or sensitivity and specificity measures, as these require converting the estimated probability outcomes into a binary regime prediction using a pre-determined (arbitrary) threshold. In the next section, we report all these performance measures for the final model selected.

¹³ The RMSE ranges between 0 and 1. See Barnston (1992) for additional discussion on this measure.

¹⁴ The C-Stat ranges from 0.5 to 1. See Hosmer and Lemeshow (2000) for further details on the interpretation of the C-Stat.

Step-wise variable selection

In order to maximize the objectiveness and unbiasedness of the final selected model, we rely on a step-wise statistical procedure to select the variables with the highest predictive power from amongst the universe of variables described in Section 2. We implement the following procedure on the 300 data partitions:

1. Consider every possible one-variable model, constructed by only including one single variable from the universe, together with the regional fixed effects and the rural indicator.¹⁵
2. Estimate each of these one-variable models and calculate their out-of-sample predictive-power metric.
3. Select the one-variable model with the highest predictive power.
4. Add one additional variable (from the remaining universe of variables) to the selected one-variable model, to construct all possible two-variable models.
5. Estimate each of these two-variable models and calculate their out-of-sample predictive power metric.
6. Select the two-variable model with the highest predictive power.
7. Add one additional variable to the selected two-variable model (from the remaining universe of variables), to construct all possible three-variable models and repeat the selection process, adding one variable at a time and selecting the one with the highest predictive power in each round, until a 12-variable model is achieved.

Additional considerations

The above provides an objective procedure to arrive at a statistically-sound model. However, in order to fulfill all of the objectives outlined in the Introduction, we allow for a degree of subjectivity in the process. Besides considering variables that are simple and easy to measure, we need to ensure that they are (1) relatively dynamic over time, (2) add value to the model, and (3) cover key dimensions around the decision to emigrate.

In terms of the first element, since one of the objectives of the index is to monitor the propensity of an individual in a household to emigrate over time, it is important that at least a subset of the selected variables likely change over time. For instance, the sex of the household head or his/her years of education would normally not change from one year to the next, while ownership of certain assets or exposure to natural shocks would. As for adding value to the model, we intend to avoid situations where two similar indicators (likely highly correlated) are jointly included. For instance, if the model already includes a variable capturing whether any of the household members in a certain age group has completed 5 or more years of education, additionally including whether any member has completed 6 or more years of education would not add real value. Finally, we must ensure that most key dimensions around the decision to emigrate are captured in the indicators proposed, including factors generally discussed in the broader migration literature and in studies specific to the Northern Triangle and Guatemala.

To achieve this, we follow the procedure described in the previous subsection to select the optimal variable at each round but we further check if the new variable selected diverts from the objectives above; if

¹⁵ The regional fixed effects and rural indicator act as shifters and are included in all step-wise regressions.

so, this new variable is excluded, and the procedure is allowed to select the second-best variable for that round.¹⁶ This process permits to have a final set of 12 variables that fulfill all the MPI objectives.

The following figure summarizes the complete methodology.

Figure 2: Outline of methodology to construct the Migration Propensity Index

4. Results

Next, we present the results from the procedure outlined above using the 48 explanatory variables available for the 2,798 households in the panel. An algorithm was constructed in Stata software to carry out the procedure and select the variable with the highest explanatory power in each round, followed by a manual inspection of the selected variable to ensure that the set of selected variables fulfilled the requirements discussed in the “Additional considerations” subsection above.

Table 1 lists the final 12 variables selected for the MPI, in the order in which they were selected based on the C-Stat metric, which is intended to maximize the model coherence. The variables that are selected using the RMSE metric are also very similar to the ones reported below, which provides additional support to the proposed model.¹⁷ Figure 3 shows, in turn, the standardized correlations between these variables and the probability to emigrate in the final model specification. These correlations measure the change in the probability to emigrate during the following 12 months (in terms of standard deviations) associated to a one standard deviation change in the variable of interest.

¹⁶ The whole estimation process is documented and is available upon request.

¹⁷ The full regression results are reported in Appendix C.

Table 1: Final variables selected for the Migration Propensity Index

#	Variable	Effect on probability to emigrate
1	Community-level emigration during the last 12 months	Positive
2	Community-level rate of migrants abroad	Positive
3	Youth aged 15-29 in household	Positive
4	Household is beneficiary of Bolsa Segura	Negative
5	Household has a migrant abroad	Positive
6	Men aged 15-29 with five years of education or above	Positive
7	Education of household head	Negative
8	Household total agricultural land is above 1000m2	Negative
9	Household incurred in home improvement in the last 12 months	Negative
10	Household owns a television	Positive
11	Grows a high-value crop	Negative
12	Age of household head	Negative
+ Rural and region indicators		(Shifters)

Figure 3: Standardized correlations between selected variables and probability to emigrate

Note: ***, ** and * denotes significance at the 1%, 5%, and 10% level.

The peer and potential network effects variables stand out as important predictors of the probability that someone in a household emigrates. Recent emigration in the community, as captured by the community-level emigration during the last 12 months, has the largest (and statistically-significant) effect, while a higher prevalence of longer-term emigrants in the community (defined as the share of households in the community who received remittances from abroad during the last 12 months) also has an important positive effect. Additionally, whether the household has an emigrant abroad (proxied by whether the household received remittances during the last 12 months) is an important pull factor positively influencing the index.

Some demographic characteristics also seem to be important decision factors. First, having any youth in the household (male or female members between 15 and 29 years of age) increases a household's MPI. This is consistent with much of the existing literature on migration, which finds youth as the main age group migrating abroad. In addition, education among this age group matters, with the propensity to emigrate increasing when at least one male member within the same age group has 5 years or more of completed education. In contrast, both education and age of the household head are negatively correlated with the decision to emigrate. Hence, having some education and skills seems to be an important trigger factor for younger household members to emigrate (most likely the sons or grandsons of the household head) but operate in the opposite direction for household heads.¹⁸

Being a beneficiary of the Bolsa Segura program also serves as an important indicator of a household having a lower propensity to emigrate, possibly signaling very poor households not having the overall means to migrate abroad. Similarly, households that report having made expenditures for home improvements during the last 12 months show a substantially lower propensity to emigrate. Finally, households with large landholdings (above 1000 m² or 0.1 hectares), those growing a high-value crop (i.e. growing a crop other than maize or beans), or those not owning a television are associated to a lower probability to emigrate, albeit showing a weaker effect than the other selected variables.

These 12 variables, together with the rural indicator and regional shifters, conform the final model. Table 2 shows several performance indicators of the model. The first column reports the full in-sample metrics; that is, the metrics using the total working sample. The second and third columns report the training in-sample and out-of-sample metrics using the 70%-30% sample partitions from the cross-validation procedure described above; the reported measures are the corresponding averages and confidence intervals across the 300 replicated subsamples.

Overall, the model shows a good predictive performance. In particular, the model's sensitivity or ability to correctly predict positive cases (i.e. households in the PPH0 sample with members who actually emigrated in the following 12 months) is in the order of 44%. In other words, this model is able, by tracking the above 12 variables alone, to correctly identify roughly one in every two households where someone ends up emigrating. At the same time, the model's specificity or ability to correctly predict negative cases (i.e. households where members do not end up emigrating) is around 95%. That is, the model correctly identifies a non-emigrating household in 19 out of 20 households.

It is worth noting that the asymmetries between the predictive power of positive cases (i.e. correctly identifying an emigrant) and negative cases (i.e. correctly identifying a non-emigrant) are determined by the low prevalence of emigrants in the population or working sample (3.5% of households reported an emigrant in our sample, which is similar to the values reported in national surveys for Guatemala such

¹⁸ Certainly, households with better-educated heads may also proxy households with a higher income that are less likely to migrate.

as the ENCOVI). When the prevalence of the modeled variable (emigration) is low, the predictive power of positive cases is lower as there are likely more false-positive outcomes due to a lower population (sample) probability; likewise, the predictive power of negative cases is higher as there are likely much less false-negative outcomes. In this line, a 44% sensitivity or positive predictive value for a 3.5% population-wise prevalence suggests a highly accurate metric (in relative terms).

Both the sensitivity and specificity metrics are very important for policymaking and targeting purposes. Correctly identifying households with future migrants is relevant for accurately targeting programs aimed at curbing emigration. Correctly identifying non-emigrating households is pertinent to better allocate program resources, which can be used in more efficient ways or to achieve objectives other than deterring emigration. In this line, the correct classification rate is a metric that combines the two measures above, reporting the percentage of cases where the model's prediction accurately identifies the actual outcomes observed. This statistic indicates that around 93% of households are correctly classified by the model as either future emigrating or non-emigrating households.

Table 2: In-sample and out-of-sample performance measures of selected model

	Full in-sample	Training in-sample		Out-of-sample	
		Mean	95% conf. interval	Mean	95% conf. interval
Root Mean Square Error	0.173	0.174	[0.165-0.185]	0.177	[0.154-0.197]
Concordance Stat.	0.864	0.866	[0.845-0.885]	0.835	[0.785-0.891]
Sensitivity	0.443	0.478	[0.406-0.541]	0.426	[0.298-0.566]
Specificity	0.948	0.945	[0.933-0.954]	0.944	[0.926-0.961]
Correct Classification Rate	0.930	0.928	[0.917-0.938]	0.925	[0.909-0.941]
Number of observations	2,713				
Number of households	1,961				
Number of migrants	97				

Note: The Root Mean Square Error (RMSE) and Concordance statistic (C-Stat) are defined in Section 2. The sensitivity (correct positive outcome rate), specificity (correct negative outcome rate) and correct classification rate (correct positive and negative outcomes rate) are calculated using a threshold of 0.1 to convert the estimated probability outcomes into a binary regime prediction and compare them with the observed outcomes.

5. Scoring

This section discusses the MPI questionnaire (module) and the scoring rules that need to be applied to each of the questions' answers in order to calculate the index. We also provide some examples on how to apply the index. As noted earlier, the idea is to have a concise, yet reliable and rigorous index that can be easily calculated in close to real time (i.e. as soon as the information is collected).

Appendix D presents the MPI questionnaire to be applied to any given household for which there is interest in determining their propensity to emigrate. It is important to remember that the resulting score is associated to the probability that at least one member within the household will emigrate during the following 12 months. Such a questionnaire can be easily implemented either in person or over the phone. After obtaining the answer to each question, the scoring instructions should be followed to obtain the

score contributed by that question (variable). Once all questions are asked, the total score should be added up and, by using the box at the bottom of the questionnaire, the analyst can determine the emigration risk category of the household. The total score can range from 0 to 100, where 100 represents the highest propensity or likelihood to emigrate.

The scores (weights) for each question are derived directly from the estimated statistical model. The logit coefficients for all variables are transformed to both contemplate the unit of measurement of that variable and that the resulting total MPI score ranges from 0 to 100, such that the weight of each variable on the emigration probability is comparable to the weight it has on the overall score.¹⁹ The scores for some of the variables, such as the household head's age and years of education, must be derived through a simple formula. If performing these calculations in the field is problematic or prone to errors, it is possible to collect the answers and process them in a centralized location. An automated excel sheet has been created for these purposes, which could easily be embedded in a tablet for field data collection.

Overall, households with final scores below 50 are considered to be of low risk in terms of their likelihood to emigrate, those with scores between 50 and 60 are considered of medium risk, and those with scores above 60 are considered of high emigration risk and should be the key candidates for targeted emigration deterrence programs. For ease of understanding, we follow a traffic-light labeling system to classify households:

- Low risk of emigration (MPI score below 50): **green household**;
- Medium risk of emigration (MPI score between 50-60): **yellow household**;
- High risk of emigration (MPI score above 60): **red household**.

It is also possible to determine a specific value for the probability to emigrate (i.e. a value between 0 and 1), as estimated from the statistical model discussed in Section 4, by using the lookup table in Appendix E. It is important to bear in mind, however, that these probabilities are centered around the observed population-level emigration rates and are thus low in absolute terms. For the purposes of the analyses above and for the scoring, we consider probabilities above 0.1 as indicative of a high propensity to emigrate.

Examples

For illustration purposes, we provide some examples on how to apply the MPI. Figure 4 shows the index score calculation for three hypothetical households with different characteristics; for example, Household A is located in the Verapaces region, while Household B is located in the Dry Corridor (Corredor Seco+Izabal), and Household C is located in the Western Highlands (Altiplano Occidental). Household A total MPI score is 32, which is below 50 and thus considered of low risk (i.e. green household) in terms of their propensity to emigrate during the following 12 months. Households B and C total MPI scores are 56 and 77, thus being classified as of medium risk (i.e. yellow household) and high risk (i.e. red household), respectively, in terms of their propensity to emigrate. Note how differences in community-level migrant rates abroad and household head's age and years of education across the different households contribute differently to the score.

Figure 5 shows (retrospective) MPI scores calculations for all available households in our panel located in Huehuetenango between the 2013 and 2014 rounds of the PPH0 surveys. Households are ordered by

¹⁹ Weights are not exactly equal to the original model since some minor rounding is done to simplify and avoid multiple decimals in the scoring system.

municipality. It is interesting to note that while there is a large heterogeneity in scores among households located across Huehuetenango, a large portion of this heterogeneity is clustered within municipalities. This is more evident in Figure 6, where households in the municipalities of San Juan Ixcoy and Concepción Huista generally exhibit a low or medium risk scoring or propensity to emigrate (with the latter showing a few high risk households) while several households in Tectitán, located very close to the Mexican border, exhibit a high risk of emigration. This is partly due to the strong influence of the peer effects (network) variables, which are shared within a community: in those communities with a large share of emigrants, the risk of future emigration is substantial.

It is pertinent to note that, if we were to apply the MPI questionnaire to the same surveyed households at present, we could also assess potential changes in their propensity to emigrate between 2013-2014 and today. This kind of exercise is relevant for the purposes of monitoring the likelihood of emigration among populations of interest over time.

Figure 4: MPI scoring examples

MPI Questions	Household A			Household B			Household C		
Region	Verapaces	→	→	Corredor Seco + Izabal	→	→	Altiplano Occidental	→	→
Household in rural area	No	→	→	Yes	→	→	Yes	→	→
Age of household head	73 years	→	→	52 years	→	→	34 years	→	→
Grows a high-value crop	No	→	→	Yes	→	→	No	→	→
Household owns TV	Yes	→	→	No	→	→	Yes	→	→
Home improvement in last 12 months	Yes	→	→	No	→	→	No	→	→
Agric. land above 1000m2	No	→	→	Yes	→	→	No	→	→
Education of household head	12 years	→	→	6 years	→	→	0 years	→	→
Men 15-29 five yrs of educ. and above	Yes	→	→	Yes	→	→	No	→	→
Household has migrant abroad	No	→	→	Yes	→	→	Yes	→	→
Household is beneficiary of Bolsa Segura	No	→	→	Yes	→	→	No	→	→
Youth aged 15-29 in household	Yes	→	→	Yes	→	→	Yes	→	→
Community-level migrants abroad	30%	→	→	40%	→	→	50%	→	→
Community-level 12-month emigration	20%	→	→	30%	→	→	50%	→	→
			LOW RISK Score = 32			MEDIUM RISK Score = 56			HIGH RISK Score = 77

Figure 5: MPI scoring for sampled households in Huehuetenango

Figure 6: MPI scoring for sampled households in selected municipalities in Huehuetenango

Lastly, the index can help at an aggregate level, to identify areas with a high, medium, or low risk of emigration, based on the prevalence of green, orange or red households in them. Continuing the previous example, if the households in the PPH0 sample were assumed to be representative of all households in the municipality, we could argue that back in 2013-2014 San Juan Ixcoy was most likely a low emigration risk (or green) municipality, Concepcion Huista a medium emigration risk (or yellow) municipality, and Tectitán a high emigration risk (or red) municipality. While this aggregation can be very useful for geographic targeting purposes, it requires collecting information from a representative sample of households in the areas of interest.

Figure 7: Hypothetical labeling of selected municipalities in Huehuetenango

6. Concluding remarks

The proposed Migration Propensity Index (MPI) is designed to provide donors, policy makers, and program implementors a tool to measure and track the likelihood that individuals in a given household will emigrate over the following 12 months. The index is based on an alternative approach that is reliable — by avoiding sensitive, direct inquiries on migration intentions—, easy to implement —by relying on a concise and simple set of questions—, and statistically-rigorous —constructed using advanced statistical methods and evaluated in terms of its out-of-sample predictive capacity. The MPI scoring system allows us to classify households as of low, medium, or high emigration risk or propensity to emigrate.

The MPI can be applied to both monitor households of interest such as direct or indirect beneficiaries of specific programs or to identify or target households at risk to emigrate. The index questionnaire is very simple and can be directly collected over the phone or embedded into a larger survey. Similarly, the non-invasive, indirect approach of the index allows its periodic application over time to the same population for monitoring purposes. Finally, if representative household data is available across broader areas (e.g.,

communities, municipalities, departments, regions), the index can be aggregated at the area-level for geographic targeting purposes, with the possibility of classifying areas as of low, medium, or high emigration potential.

We calibrate the index for the case of Guatemala, but the MPI could be easily adapted and extended to other neighboring countries such as El Salvador and Honduras. The index should also be regarded as a dynamic instrument, developed for predictive purposes using the best currently available data, but whose precision can be continuously validated and improved as more data is collected. At present, the proposed model correctly identifies 93% of eventual emigrating and non-emigrating households, with its precision prone to further refinement, especially in identifying eventual emigrants, as more panel data becomes available.

REFERENCES

- Adams, R.H.J. and Cuecuecha, A., 2010. Remittances, Household Expenditure and Investment in Guatemala. *World Development* 38(11): 1626-1641.
- Barnston, A., 1992. Correspondence Among the Correlation [Root Mean Square Error] and Heidke Verification Measures; Refinement of the Heidke Score. *Weather and Forecasting* 7: 699-709.
- Cohn, D., Passel, J., and Gonzalez-Barrera, A., 2017. Rise in U.S. Immigrants From El Salvador, Guatemala and Honduras Outpaces Growth from Elsewhere. Pew Research Center, December 7, <https://www.pewresearch.org/hispanic/2017/12/07/rise-in-u-s-immigrants-from-el-salvador-guatemala-and-honduras-outpaces-growth-from-elsewhere/> (accessed March 2020).
- Congressional Research Service, 2019a. Recent Migration to the United States from Central America: Frequently Asked Questions. CRS REPORT R45489, January 29, <https://fas.org/sgp/crs/row/R45489.pdf> (accessed March 2020).
- Congressional Research Service, 2019b. Central American Migration: Root Causes and U.S. Policy. IN FOCUS IF11151, June 13, <https://fas.org/sgp/crs/row/IF11151.pdf> (accessed March 2020).
- Creatives Associates International, 2019. Why Migrants Risk it All: What Drives Unauthorized Migration to the U.S.? What Shapes the Decision to Migrate? Mimeo, July.
- de Brauw, A., 2019. Determinants of Migration Among Rural Youth Throughout the World. IFPRI Discussion Paper 1898.
- De Haan, A., 1999. Livelihoods and Poverty: The Role of Migration - A Critical Review of the Migration Literature. *The Journal of Development Studies* 36(2): 1-47.
- Docquier, F. and Marfouk, A., 2006. International Migration by Education Attainment, 1990–2000. In *International Migration, Remittances and the Brain Drain*. Washington DC: World Bank, 151-199.
- Donnelly, R. and Hagan, J.M., 2014. The Dangerous Journey: Migrant Smuggling from Mexico and Central America, Asia, and the Caribbean. In *Hidden Lives and Human Rights in the United States: Understanding the Controversies and Tragedies of Undocumented Immigration*, 71-106.
- Food and Agriculture Organization – FAO, 2018. The State of Food and Agriculture 2018: Migration, Agriculture and Rural Development. FAO, Rome.
- Galor, O. and Stark, O., 1991. The Probability of Return Migration, Migrants' Work Effort, and Migrants' Performance. *Journal of Development Economics* 35(2): 399-405.
- Greene, W.H., 1993. *Econometric Analysis*. New York: Macmillan Publishing Company.
- Hosmer, D.W., and Lemeshow, S., 2000. *Applied Logistic Regression* (2nd Edition). John Wiley & Sons.
- Instituto Nacional de Estadística – INE, 2018. Herramientas de Datos del Censo 2018, <https://www.censopoblacion.gt/> (accessed November 2019).
- International Food Policy Research Institute – IFPRI, 2016. Informe de Evaluación de Impacto del Plan del Pacto Hambre Cero. Mimeo
- International Fund for Agricultural Development – IFAD, 2016. Rural Development Report 2016. *Fostering inclusive rural transformation*. Rome.
- Isacson, A., Meyer, M. and Morales, G., 2014. Mexico's Other Border: Security, Migration, and the Humanitarian Crisis at the Line with Central America. Washington Office on Latin America, 2-44.
- Jessoe, K., Manning, D.T., and Taylor, J.E., 2018. Climate Change and Labour Allocation in Rural Mexico: Evidence from Annual Fluctuations in Weather. *The Economic Journal* 128(608): 230-261.
- Mahajan, P., and Yang, D., 2020. Taken by Storm: Hurricanes, Migrant Networks, and US Immigration. *American Economic Journal: Applied Economics* 12(2): 250-277.
- Mattoo, A., Neagu, I.C. and Özden, C., 2005. Brain Waste. Educated Immigrants in the US Labor Market, World Bank Policy Research Paper 3581.
- McKenzie, D., 2006. Beyond Remittances: The Effects of Migration on Mexican Households. In *International Migration, Remittances and the Brain Drain*. Washington DC: World Bank, 123-147.

- Munshi, K., 2003. Networks in the Modern Economy: Mexican Migrants in the U.S. Labor Market. *The Quarterly Journal of Economics* 118(2): 549-599.
- National Immigration Forum, 2019. Push or Pull Factors: What Drives Central American Migrants to the U.S.?, https://immigrationforum.org/wp-content/uploads/2019/07/PushPull-Factors_Paper_Final.pdf (accessed March 2020).
- Rubenstein, J.M., 2017. *The Cultural Landscape: An Introduction to Human Geography* (12th Edition). Chapter 3: Migration. Pearson.
- Schreiner, M., 2013. Simple Poverty Scorecard Poverty-Assessment Tool: Bangladesh. Mimeo, Microfinance Risk Management.
- Singh, S., 2015. Increased Burden on Women and Male Out-Migration: An Analysis of Khul Gad Micro Watershed of the Kumoun Himalaya. *International Journal of Social Work and Human Services Practice* 3(3): 118-123.
- Skeldon, R., 1997. *Migration and Development: A Global Perspective*. London: Routledge.
- Stark, O., 1991. *The Migration of Labor*. Oxford, UK: Basil Blackwell.
- Torres, J.M., Alcántara, C., Rudolph, K.E. and Viruell-Fuentes, E.A., 2016. Cross-Border Ties as Sources of Risk and Resilience: Do Cross-Border Ties Moderate the Relationship Between Migration-Related Stress and Psychological Distress for Latino Migrants in the United States? *Journal of Health and Social Behavior* 57(4): 436-452.
- Torres, J.M. and Wallace, S.P., 2013. Migration Circumstances, Psychological Distress, and Self-Rated Physical Health for Latino Immigrants in the United States. *American Journal of Public Health* 103(9): 1619-1627.
- United Nations, 2019. International Migrant Stock 2019. Department of Economic and Social Affairs. United Nations Database, <https://www.un.org/en/development/desa/population/migration/data/estimates2/estimates19.asp> (accessed March 2020).
- Usher, D., 1977. Public Property and the Effects of Migration upon other Residents of the Migrants' Countries of Origin and Destination. *Journal of Political Economy* 85(5): 1001-1020.
- Valentine, J.L., Barham, B., Gitter, S., and Nobles, J., 2017. Migration and the Pursuit of Education in Southern Mexico. *Comparative Education Review* 61(1): 000.
- World Bank, 2018. *Moving for Prosperity: Global Migration and Labor Markets*. Policy Research Report. The World Bank Group.
- Yang, D. and Martinez, C., 2006. Remittances and Poverty in Migrants' Home Areas: Evidence from the Philippines. In *International Migration, Remittances and the Brain Drain*. Washington DC: World Bank, 81-121.

APPENDIX

Appendix A: Sampling design of panel survey

The Plan del Pacto Hambre Cero (PPH0) focused on the 176 municipalities with the highest chronic malnutrition (stunting) rates in the country. Yet, the implementation of PPH0 was planned to occur gradually over time during three separate phases: a first phase focusing on the 33 municipalities with the highest prevalence of stunting in the country; a second phase focusing on the next 50 municipalities with the highest stunting prevalence; and a third phase including the remaining 83 municipalities with a high prevalence of chronic malnutrition (greater than the country's median). The corresponding cutoff stunting rates for each phase were 42.9%, 60%, and 69.8%. Based on this gradual implementation of the PPH0, the ten closest municipalities above and below each of these three cutoff rates were selected as the sampling frame for the panel dataset, resulting in a total of 60 municipalities.²⁰

Once the sampling frame was defined, the households selected for the panel survey had to include at least one member belonging to one of the three population groups targeted by the PPH0; that is, the household had to include children under 5 years of age, pregnant women, or women of childbearing age (between 15 and 49 years old). The final selection of the sample of households was carried out in two stages:

1. In a first stage, 180 Primary Sampling Units (PSUs) were selected probabilistically (with probability proportional to size, in terms of number of dwellings), using the 2002 Census tracks in the selected municipalities, as defined by the National Statistical Institute (INE). The selection was stratified across municipalities such that 3 PSUs were selected from each of them. The map below shows the selected PSUs across the country.

Figure A.1: PSUs selected for the panel survey

²⁰ This sampling frame selection was linked to the intended impact evaluation strategy of the program (a regression discontinuity design). See IFPRI (2016) for additional details on the evaluation of the PPH0.

2. During the second stage, a full listing was first conducted to identify all households with eligible individuals within each chosen PSU. From this sample frame, households with at least one eligible individual were then selected probabilistically according to the criteria presented in the table below. The sampling was done progressively: first pregnant women, then children under one year old, and so forth; excluding a household from the list after being selected. Each selection was made with equal probability among all eligible individuals in the remaining households. As a result, 14 households per PSU were selected in total. These households were the ones surveyed.

Table A.1: Number of sampled households per PSU

Eligible individual		Number of sampled households
1	Pregnant women	3
2	Children between 0 and 0.99 years old	3
3	Children between 1 and 1.99 years old	2
4	Children between 2 and 2.99 years old	2
5	Children between 3 and 3.99 years old	2
6	Children between 4 and 4.99 years old	2
Total		14

Appendix B: Universe of variables considered for the model selection

#	Name	Type	Definition	Timing
1	Gender of household head	Binary	One if the household head is male; zero if female	At time of the interview
2	Age of household head	Integer	Age of the household head in years	At time of the interview
3	Education of household head	Integer	Number of completed years of education of the household head	At time of the interview
4	Indigenous household	Binary	One if the interviewee considers him or herself to be indigenous; zero otherwise	At time of the interview
5	Household size	Integer	Number of household members	At time of the interview
6	Men aged 15-29 in household	Binary	One if the household contains at least one male member aged 15-29 years; zero otherwise	At time of the interview
7	Women aged 15-29 in household	Binary	One if the household contains at least one female member aged 15-29 years; zero otherwise	At time of the interview
8	Youth aged 15-29 in household	Binary	One if the household contains at least one (male or female) member aged 15-29 years; zero otherwise	At time of the interview
9	Men aged 30 or above in household	Binary	One if the household contains at least one male member aged 30 years or above; zero otherwise	At time of the interview
10	Women aged 30 or above in household	Binary	One if the household contains at least one female member aged 30 years or above; zero otherwise	At time of the interview
11	Adults aged 30 or above in household	Binary	One if the household contains at least one (male or female) member aged 30 years or above; zero otherwise	At time of the interview
12	Men aged 15-29 with five years of education or above	Binary	One if the household contains at least one male member aged 15-29 with five or more complete years of education; zero otherwise	At time of the interview
13	Women aged 15-29 with five years of education or above	Binary	One if the household contains at least one female member aged 15-29 with five or more complete years of education; zero otherwise	At time of the interview
14	Men aged 15-29 with primary education or above	Binary	One if the household contains at least one male member aged 15-29 with completed primary education or above; zero otherwise	At time of the interview
15	Women aged 15-29 with primary education or above	Binary	One if the household contains at least one female member aged 15-29 with completed primary education or above; zero otherwise	At time of the interview
16	Men aged 15-29 with 1 year of secondary education or above	Binary	One if the household contains at least one male member aged 15-29 with one or more complete years of secondary education; zero otherwise	At time of the interview
17	Women aged 15-29 with 1 year of secondary education or above	Binary	One if the household contains at least one female member aged 15-29 with one or more complete years of secondary education; zero otherwise	At time of the interview
18	Dwelling has finished walls	Binary	One if the dwelling walls are made of block or brick; zero otherwise	At time of the interview
19	Dwelling has finished ceiling	Binary	One if the dwelling ceiling is made of concrete or tiles; zero otherwise	At time of the interview
20	Dwelling has finished floors	Binary	One if the dwelling floors are made of cement, brick, ceramics, polished wood, or boards; zero otherwise	At time of the interview
21	Dwelling connected to water	Binary	One if the household obtains water for consumption from pipes in their dwelling or land; zero otherwise	At time of the interview
22	Dwelling connected to electricity	Binary	One if the dwelling is connected to an electric supply network; zero otherwise	At time of the interview
23	Dwelling connected to sewerage	Binary	One if the dwelling has a toilet connected to sewer of exclusive use for the household; zero otherwise	At time of the interview
24	Household total agricultural land	Continuous	Total hectares of agricultural land worked by a household, 0 if none	During the last 12 months
25	Household has agricultural land	Binary	One if the household works any agricultural land; zero otherwise	During the last 12 months

#	Name	Type	Definition	Timing
26	Household total agricultural land is above 1000m2	Binary	One if the household works agricultural land of 1000 square meters or above; zero otherwise	During the last 12 months
27	Produces a subsistence crop	Binary	One if the household cultivates maize or beans; zero otherwise	During the last 12 months
28	Produces a high-value crop	Binary	One if the household cultivates coffee, cardamom, a horticultural crop, or other crops apart from maize and beans; zero otherwise	During the last 12 months
29	Household owns a cellphone	Binary	One if the household owns a cellphone; zero otherwise	At time of the interview
30	Household owns a television	Binary	One if the household owns a television; zero otherwise	At time of the interview
31	Household owns a radio	Binary	One if the household owns a radio; zero otherwise	At time of the interview
32	Household owns a fridge	Binary	One if the household owns a fridge; zero otherwise	At time of the interview
33	Household owns a computer	Binary	One if the household owns a computer; zero otherwise	At time of the interview
34	Household has cable television	Binary	One if the household has a cable television subscription; zero otherwise	At time of the interview
35	Household owns houses or land	Binary	One if the household owns any real estate property; zero otherwise	At time of the interview
36	Household owns a vehicle	Binary	One if the household owns a vehicle; zero otherwise	At time of the interview
37	Household owns agricultural machinery or equipment	Binary	One if the household owns agricultural machinery or equipment; zero otherwise	At time of the interview
38	Household owns livestock	Binary	One if the household owns livestock; zero otherwise	At time of the interview
39	Household purchased durable goods during last 12 months	Binary	One if the household made any expenditures on new durable goods during the last 12 months; zero otherwise	During the last 12 months
40	Household purchased furniture during last 12 months	Binary	One if the household made any expenditures on new furniture during the last 12 months; zero otherwise	During the last 12 months
41	Household purchased vehicles during last 12 months	Binary	One if the household made any expenditures on new vehicles during the last 12 months; zero otherwise	During the last 12 months
42	Household made home improvement during last 12 months	Binary	One if the household made any expenditures on construction, expansion, or improvement of their dwelling during the last 12 months; zero otherwise	During the last 12 months
43	Household is beneficiary of Bolsa Segura	Binary	One if household is beneficiary of social program Bolsa Segura; zero otherwise	During the last 12 months
44	Household is beneficiary of Bono Seguro de Salud	Binary	One if household is beneficiary of social program Bono Seguro de Salud; zero otherwise	During the last 12 months
45	Household affected by a natural disaster	Binary	One if the household has been affected by a natural disaster (excess rain, flood, earthquake, volcano, pests, drought, frost,) during the last 3 years; zero otherwise	During the last 3 years
46	Household has a migrant abroad	Binary	One if the household received remittances during the last 12 months; zero otherwise	During the last 12 months
47	Community-level emigration during the last 12 months	Fraction [0-1]	Share of households in a household's community in which at least one member emigrated during the last 12 months	During the last 12 months
48	Community-level rate of migrants abroad	Fraction [0-1]	Share of households in a household's community that received remittances during the last 12 months	During the last 12 months

Appendix C: Full Sample Regression Results – Marginal Effects

	(1) Will migrate in next 12 months
Community-level migration during the last 12 months	0.116*** (0.014)
Community-level rate of migrants abroad	0.051 (0.032)
Youth aged 15-29 in household	0.017 (0.015)
Household is beneficiary of Bolsa Segura	-0.034* (0.020)
Household has a migrant abroad	0.022*** (0.009)
Men aged 15-29 with five years of education or above	0.018** (0.007)
Education of household head	-0.003** (0.001)
Household total agricultural land is above 1000m2	-0.007 (0.007)
Household incurred in home improvement in the last 12 months	-0.033 (0.023)
Household owns a television	0.007 (0.007)
Grows a high-value crop	-0.004 (0.008)
Age of household head	-0.000 (0.000)
Located in rural area	0.029** (0.012)
Region: Centro + Pacifico	0.011* (0.006)
Region: Altiplano Occ	0.043*** (0.005)
Region: Corredor Seco + Izabal	0.033** (0.014)
Region: Verapaces	0.017** (0.008)
Number of observations	2,713

Note: This table shows the full-model results from a logistic specification of an indicator variable capturing whether at least one individual in a given household migrated within 12 months since the date of the interview and selected explanatory variables that maximize the predictive power of the model, as described in the text. The coefficients are average marginal effects across the entire sample. Standard errors, clustered at the household level, are in parentheses. ***, **, and * denote statistical significance at the 1%, 5%, and 10% level, respectively.

Appendix D: Migration Propensity Index questionnaire

Question	Answer	Instruction	Score
a. Approximately how many households live in your community?	_____ households		
b. In how many of these households did someone migrate abroad during the last 12 months?	_____ households	Calculate: (b / a) x 26	
c. How many of these households currently receive remittances?	_____ households	Calculate: (c / a) x 11	
d. Does someone in your household currently receive remittances?	No Yes	0 5	
e. How old is the head of your household?	_____ years	Calculate: (95 – e) x 0.05	
f. How many years of completed education does the head of your household have?	_____ years	Calculate: (17 – f) x 0.65	
g. Are there any men or women between 15 and 29 years old living in your household?	No Yes	0 4	
h. Out of all the men between 15 and 29 years old in your household, has anyone completed 5 or more years of education?	No Yes	0 4	
i. Is your household a beneficiary of Bolsa Segura?	No Yes	7 0	
j. Does anyone in your household own more than 1,000 m ² (0.1 hectares) of agricultural land?	No Yes	2 0	
k. Has your household incurred in any home improvement or expansion expenses during the last 12 months?	No Yes	7 0	
l. Does your household own a TV?	No Yes	0 2	
m. Does anyone in your household cultivate agricultural crops other than maize or beans?	No Yes	1 0	
n. Is the household located in a rural area?	No Yes	0 6	
o. In which region is the household located?	Centro + Pacifico	0	
	Verapaces	3	
	Corredor Seco + Izabal	8	
	Altiplano Occidental	10	

Total score:

If score is... ...then

Under 50	Low risk
Between 50 and 60	Medium risk
Above 60	High risk

Appendix E: MPI score-to-probability lookup table

When score is... ...above:	and below:	...the probability to emigrate is:	Risk category:
10	12	0.0%	LOW RISK
12	14	0.0%	
14	16	0.0%	
16	18	0.0%	
18	20	0.0%	
20	22	0.0%	
22	24	0.0%	
24	26	0.1%	
26	28	0.1%	
28	30	0.1%	
30	32	0.2%	
32	34	0.2%	
34	36	0.3%	
36	38	0.4%	
38	40	0.5%	
40	42	0.7%	
42	44	1.0%	
44	46	1.3%	
46	48	1.8%	
48	50	2.4%	
50	52	3.2%	MEDIUM RISK
52	54	4.3%	
54	56	5.8%	
56	58	7.7%	
58	60	10.2%	
60	62	13.5%	HIGH RISK
62	64	17.5%	
64	66	22.3%	
66	68	28.2%	
68	70	34.7%	
70	72	41.9%	
72	74	49.8%	
74	76	57.5%	
76	78	65.3%	
78	80	71.7%	
80	82	77.3%	
82	84	82.4%	
84	86	86.7%	
86	88	89.5%	
88	90	91.6%	
90	92	94.3%	
92	94	95.4%	

ALL IFPRI DISCUSSION PAPERS

All discussion papers are available [here](#)

They can be downloaded free of charge

INTERNATIONAL FOOD POLICY RESEARCH INSTITUTE

www.ifpri.org

IFPRI HEADQUARTERS

1201 Eye Street, NW
Washington, DC 20005 USA
Tel.: +1-202-862-5600
Fax: +1-202-862-5606
Email: ifpri@cgiar.org