New York, Abu Dhabi, London or Stay at Home? Using a Cross-Nested Logit Model to Identify Complex Substitution Patterns in Migration

Michel Beine [*] Michel Bierlaire [†] Frédéric Docquier [‡]

January, 30, 2021

Report TRANSP-OR 210130
Transport and Mobility Laboratory
School of Architecture, Civil and Environmental Engineering
Ecole Polytechnique Fédérale de Lausanne
transp-or.epfl.ch

[*] University of Luxembourg, IZA and CES-Ifo, michel.beine@uni.lu

[†] École Polytechnique Fédérale de Lausanne (EPFL), School of Architecture, Civil and Environmental Engineering (ENAC), Transport and Mobility Laboratory, Switzerland, michel.bierlaire@epfl.ch

[‡] Luxembourg Institute of Socio-Economic Research, FERDI, CReAM, and IZA, frederic.docquier@liser.lu

Abstract

The question of how people revise their decisions about whether to emigrate, and where to, when facing changes in the global environment is of critical importance in migration literature. We propose a cross-nested logit (CNL) approach to generalize the way deviations from the IIA (independence from irrelevant alternatives)) hypothesis can be tested and exploited in migration studies. Compared with the widely used logit model, the structure of a CNL model allows for more sophisticated substitution patterns between destinations. To illustrate the relevance of our approach, we provide a case study using migration aspiration data from India. We demonstrate that the CNL approach outperforms standard competing approaches in terms of quality of fit, has stronger predictive power, implies stronger heterogeneity in responses to shocks, and highlights complex and intuitive substitution patterns between all possible alternatives. In particular, we shed light on the low degree of substitutability between the home and foreign alternatives as well as on the subgroups of countries that are considered by potential Indian movers as highly or poorly substitutable.

JEL Classification: C25, F22, J61.

Keywords: International migration; Discrete choice modelling; Independence from irrelevant alternatives; Cross-nested logit; Migration aspirations.

1 Introduction

International migration is at the forefront of policy debates in most countries around the world. In industrialized nations, the proportion of foreigners in total population increased from 4.5 to 12 percent between 1960 and 2019, stirring up fears about economic costs for natives, loss of national identity, and integration issues. In poor countries, international migration raises concerns about the brain drain of highly-skilled workers, as college and university graduates have a much greater propensity to emigrate internationally than the less educated. Hence, the questions of how many people migrate (i.e., migration intensity), which people migrate first or are more likely to migrate (i.e., migrants' selection), and where migrants choose to settle (i.e., migrants' sorting) have been analyzed from all possible angles in recent literature. Specifically, understanding how people revise their decisions about whether to emigrate, and where to, when facing changes in the global environment is of crucial importance for decision-makers. We propose a new approach that address such fundamental issues.

The overwhelming majority of previous studies rely explicitly or implicitly on the logit approach, which implies the validity of the property of independence from irrelevant alternatives (IIA) at the disaggregate level and specific patterns in the substitution between potential locations. IIA implies that cross elasticities due to a change in one destination's attributes are identical for all alternatives. As argued below, this substitution pattern might be too restrictive in migration choice settings, as it is very unlikely that all destinations can be reasonably treated as equally substitutable. Some destinations share common unobserved features, which make them more similar compared to others.[1] While deviations from IIA – leading to what are known as multilateral resistances – have been extensively examined in the trade literature (Anderson and van Wincoop, 2003, Anderson and van Wincoop, 2004), very few contributions have dealt in depth with this issue in migration studies. Some recent studies account for unobserved differences between home and foreign destinations, assuming that the IIA assumption holds across foreign destinations (Ortega and Peri, 2013, Buggle et al., 2019, Monras, 2020). A few other innovative works account for a violation of IIA across foreign destinations, but use a somewhat arbitrary partitioning of countries (Bertoli and Fernández-Huertas Moraga, 2015, Bredtmann et al., 2017).

We generalize the way deviations from the IIA hypothesis can be tested and exploited in migration studies by using a cross-nested logit (CNL) modelling approach relying on an overlapping nest structure. The CNL model was introduced in the late 1990's (Vovsha, 1997, Ben-Akiva and Bierlaire, 1999) and is an extension of the popular nested logit (NL) approach (Ben-Akiva, 1973). An appealing feature of the CNL model is that it provides a highly intuitive and flexible way of partitioning the choice set. The NL model is based on a strict partitioning of the choice set into "nests." The substitutability is then stronger among destinations within the same nest than between destinations belonging to two different nests. The CNL model relaxes the requirement to have a one-dimensional partition, and allows each alternative to belong to two different nests. This allows for more complex substitution patterns. Moreover, as the NL is a restriction of the CNL, standard likelihood ratio tests can be used to decide which model is most appropriate in a particular context.

The CNL approach has been used in different fields, such as the modeling of transportation choices. For the first time, we implement it in the context of international migration. In line with Ortega and Peri, 2013, Buggle et al., 2019 and Monras, 2020, our proposed specification first distinguishes between the home and foreign destinations in order to capture unobserved differences between the two types of location – or between "stayers" and "movers." In addition, we also identify a set of overlapping subgroups of foreign destinations defined along several dimensions that are usually perceived as relevant in migration literature (e.g., level of economic development, quality of institutions, geographic location, language spoken, and contiguity). The empirical estimation of the CNL model allows us to quantify the degree of similarity between destinations within each nest, and the complex substitution patterns between all possible locations resulting from the overlapping nest structure.

We provide a case study that illustrates the relevance of our approach. We use microdata on migration aspirations from India over the period 2007-2016. Data is taken from the Gallup World Poll (GWP) surveys that document individuals' willingness

[1] For instance, European countries might share common features such as norms, values and cultural traits that make them more similar to each other in the view of prospective migrants compared with non European destinations. Persian Gulf countries are more similar to each other in terms of culture, religion and religiosity, and gender-egalitarian views as well as migration policies.

to emigrate – if they had the opportunity to do so – as well as aspiring migrants' preferred destination. We find strong evidence that the CNL approach performs better than standard competing approaches such as a logit or a minimal NL model with two nests. First, likelihood ratio tests show that the CNL provides a better fit to the data. Second, our validation experiments show that CNL provides superior out-of-sample predictions. Third, CNL delivers different values of the estimated elasticities involving the main determinants such as income at destination. Fourth, CNL generates significantly different substitution patterns across destinations. To illustrate this, we simulate a counterfactual scenario in which the U.S. is made inaccessible to all Indian respondents. While the logit and the NL models imply almost identical substitution across foreign locations (close to proportionate shifting), the CNL generates much richer substitution patterns, translating into large differences in migration responses across alternatives. Our case study sheds light on the low degree of substitutability between the home and foreign alternatives, as well as on the subgroups of countries that are considered by Indian potential movers as highly or poorly substitutable for the U.S.

We contribute to a large and growing body of literature on the identification of factors affecting the decision to move and the destination choice conditional on wishing to leave. Existing literature has focused on regular migrants (Mayda, 2010, Grogger and Hanson, 2011, Beine et al., 2011, McKenzie and Rapoport, 2011, Ortega and Peri, 2013, Cattaneo and Peri, 2016, Dao et al., 2018), on asylum seekers and refugees (Bertoli et al., 2016, Bertoli, Bruecker and Fernández-Huertas Moraga, 2020, Dustmann et al., 2019, Hatton, 2016, Hatton, 2017, Hatton, 2020, Beine et al., 2021), or on undocumented migrants (Bazzi et al., 2018, Friebel et al., 2019, Gathmann, 2007, Jandl, 2007). The main bulk of the literature has examined the determinants of aggregate flows or stocks of international migrants using gravity-like models. It has identified major determinants of international flows such as income disparities, differences in institutional quality, geographic, linguistic, and cultural distance, migrant networks, and changes in migration laws and policies as well as push factors such as conflicts and climatic shocks.

Other studies have used microdata to compare the characteristics of households that are not directly exposed to migration, with households that have a family member abroad (e.g., Foster and Rosenzweig, 2002, Chort and Senne, 2015, Chort and Senne, 2018). The literature relying on microdata is less developed, as access to information about individual cross-border movements across a wide range of destinations is much less widespread. From the perspective of surveys and national censuses, emigrants are demographically similar to deaths, in that they cannot be interviewed. In addition, when family members who are left behind are asked questions about relatives abroad, existing micro studies suffer from severe attrition issues among migrant households due to the fact that the remaining members are likely to have joined another household (Bertoli and Murard, 2020). For this reason, other micro studies have focused on migration aspirations, comparing individuals who want to stay in their home country with those who have not yet migrated but express a desire to move (Manchin and Orazbayev, 2014, Dustmann and Okatenko, 2014, Docquier et al., 2015, Manchin and Orazbayev, 2018, Bertoli and Ruyssen, 2018, Ruyssen and Salomone, 2018, Beine et al., 2020, Docquier et al., 2019). The advantage of using data on migration aspirations (such as GWP data) is that it can be used to predict future migration pressures (Docquier et al., 2014, Bertoli and Ruyssen, 2018) and is less influenced by out-selection factors such as migration laws and policies (Özden et al., 2018). Further, such data better captures the factors governing individuals' motivation to move. It also helps identify the self-selection factors of international migration and the substitution patterns within the choice set from the point of view of potential migrants.

In most existing studies, a discrete location-choice problem for individuals is usually derived from a Random Utility Maximization (RUM) problem, where the indirect utility of choosing a particular location is expressed as the sum of a dyad-specific deterministic component and a dyad-person-specific random taste shock. The deterministic component combines the mean levels of benefits and moving costs associated with a particular dyad of countries and skill level. The random term represents the unobservable determinants that enter the utility function and are independent of the deterministic component. Assuming that random taste shocks are independent (i.e., uncorrelated across destinations) and are identically and extreme-value distributed implies that optimal location decisions satisfy the property of independence from irrelevant alternatives (IIA). IIA implies that cross-elasticities are the same across all pairs of potential destinations. With micro data, the IIA hypothesis results in the popular logit model (ML), implying that the relative probability of choosing between two alternative options depends purely on the attractiveness of these two options (McFadden, 1973). With macro data on migration flows or stocks, IIA provides the state-of-the-art microfoundations for gravity-like models of migration, implying that the ratio of dyadic migrants to stayers only depends on the characteristics of the relevant dyad (Beine et al., 2016).

Sources of violation from the IIA hypothesis stem from modeling imperfections (i.e., the existence of unobserved location characteristics or individual traits and preferences that are correlated across destinations) (Bertoli and Fernández-Huertas Moraga, 2013, Bertoli and Fernández-Huertas Moraga, 2015, Bredtmann et al., 2017), from individuals' rational decisions to limit the costly acquisition of information to a subset of alternatives (Bertoli, Fernández-Huertas Moraga and Guichard, 2020), from the fact that some (ex-post) individual characteristics might have been acquired after moving to a place of residence (de la Croix et al., 2020), or from general equilibrium and spillover effects (decisions made by a group of individuals affecting the distribution of attributes). Failing to capture these effects can result in correlated random taste shocks and thus calls for the use of alternative modeling approaches.

A first set of studies that account for deviations from IIA distinguishes between home and foreign destinations or equivalently, between stayers and movers. Although they usually represent the largest group of the population and are key to elicit factors explaining the likelihood of emigrating, stayers are not always accounted for in empirical migration studies. Stayers, however, are likely to differ from movers in terms of their unobserved characteristics and preferences.[2] The imperfect substitution between home and foreign destinations has been formalized in a small number of studies. Ortega and Peri, 2013 developed a micro-founded gravity approach in which the stochastic component of the movers shares a common component across foreign destinations.[3] More recently, Buggle et al., 2019 used individual data to identify factors influencing the emigration decisions and destination choices of Jewish refugees during the 1930s in Nazi Germany. They developed an explicit nested logit approach along the lines of McFadden, 1978, allowing them to identify the factors inducing Jewish individuals to leave Germany.[4] In the same vein, Monras, 2020 also developed a recent analysis of international migration allowing for a distinction between the home location and foreign alternatives.

Although they separate home countries and foreign ones, the studies above assume that all foreign destinations are equally substitutable (i.e., that the IIA hypothesis holds across foreign alternatives). Few attempts have been made to account for heterogeneous substitutability across subsets of destinations. One important exception is by Bertoli and Fernández-Huertas Moraga, 2013, who linked the concept of multilateral resistance to migration to deviation from the IIA property in gravity models of migration.[5] Another notable exception by the same authors (Bertoli and Fernández-Huertas Moraga, 2015) introduced a subset of foreign destinations, allocated in non-overlapping nests, in a micro-founded gravity model to study the impact of bilateral immigration restrictions on migration flows. They show that their approach delivers different estimates of the effect of these restrictions compared with a specification derived from the usual logit model.[6] More recently, Bredtmann et al., 2017 estimated a Random Parameter Model with country dummies capturing nests of similar regions of the same country to evaluate the sensitivity of their estimates to the possible rejection of the IIA hypoth-

[2] At the world level, international migrants only represent 3.5% of the population (Dao et al., 2018), and college-educated migrants represent 5.5% of the high-skilled population (Docquier and Rapoport, 2012). When focusing on migration aspirations from the GWP data, intended migrants account for about 20% of the respondents.

[3] For the purposes, the working equation includes origin-time fixed effects and allows for the estimation of the effects of dyadic and destination specific factors. The parameter of dissimilarity between foreign destinations remains unidentified.

[4] In particular, they estimated a two-level nested-logit approach allowing the inclusion of all potential foreign destinations within a separate nest from the "stay" alternative. In the bottom part of the model, they nevertheless assume IIA between foreign destinations.

[5] Bertoli and Fernández-Huertas Moraga, 2013 show that the concept of multilateral resistance to migration in gravity models is intrinsically related to the fact that the underlying stochastic component of the utilities follows a GEV generation function corresponding to the CNL, similar to the one we use in this paper. They show that failure to account for this (such as in standard gravity models relying on the EVT of type 1 distribution) results in spatially correlated error terms and, more importantly, in endogeneity issues. While they did not estimate the structural parameters of the CNL, they used the Common Correlated Effects estimator of Pesaran, 2006 to correct for these issues.

[6] The approach within the traditional gravity model is obtained by introducing origin-nest fixed effects. This approach does not allow to identify the parameters capturing the dissimilarity of the correlation of destination within the same nest. Furthermore, the dissimilarity parameters are assumed to be the same across the different nests.

esis. Their results do not empirically support the existence of these nests and therefore tend therefore to validate the use of a logit approach. By contrast, our CNL approach evidences very heterogeneous substitution patterns across subsets of destinations and clearly does not support the logit or the two-nest NL frameworks.

The paper is organized as follows. Section 2 details our approach to modelling international migration decisions. Section 3 provides details about the data we use in the model, presents the estimations results, documents the performance of our modelling approach, and discusses its implications for cross-destination substitution patterns. Section 4 concludes.

2 A CNL Model for International Migration

We model individual n's decisions about whether to emigrate, and where to, from a given country of origin. The set of potential locations j $(j = 0, 1, ..., J)$ includes the domestic location indexed by 0 (the alternative choice of individuals not willing to leave their country, referred to as stayers) as well as the J foreign locations (the alternatives chosen by individuals willing to leave their country, referred to as movers). In the case study in Section 3, we focus on migration aspiration data between 2007 and 2016 for Indian individuals aged 15 and over. Our sample consists of pooled cross-sectional data, as individuals are not followed over time. The key variable of interest is $P_n(j|C_n)$, representing the probability that individual n is willing to locate in destination j that belongs to the choice set C_n. Since individuals are rarely constrained in terms of location choice, the choice set can be assumed to be identical for all individuals (i.e., $C_n = C$ $\forall n$).

In line with the RUM approach, individuals maximize their utility over all possible destinations. Formally, the utility of individual n of choosing destination j is expressed as U_{jn} and can be additively decomposed into a deterministic component V_{jn} and a stochastic component ε_{jn}:

$$U_{jn} = V_{jn} + \varepsilon_{jn}. \tag{1}$$

We discuss below the assumptions about ε_{jn} and the specification of V_{jn}.

2.1 Stochastic Component of Utility

Many studies in relevant literature assume that ε_{jn} is independent and identically distributed across destinations and individuals, and follows an Extreme Value Distribution (EVD) of type 1. This is the underlying assumption of the traditional logit model (McFadden, 1973). While mathematically convenient, this assumption is violated in most contexts where discrete choice models are applied (Train, 2009). We claim that the location choice is no exception in this regard. As stated above, correlation across some subsets of destinations is a natural ingredient of location decisions for several reasons. First, intended stayers and intended movers are very different, and foreign destinations are therefore likely to be more correlated with each other than with the domestic destination. This has motivated the use of separate nests for the domestic location and the foreign potential locations in recent studies (Buggle et al., 2019, Monras, 2020). Second, some foreign destinations will be more correlated among themselves compared with others. While careful specification of the deterministic component V_{jn} might capture some part of these correlation patterns, unobserved shared characteristics will result in correlation in the stochastic terms. Hence, it is unlikely that the ε_{jn}s comply with the independence assumption. Third, random terms are likely to be spatially correlated if migrants rationally decide to limit the acquisition of information to a subset of alternatives (Bertoli, Fernández-Huertas Moraga and Guichard, 2020), if some of their (ex-post) observed characteristics have been acquired after moving (de la Croix et al., 2020), or if migration influences the distribution of country characteristics.

We adopt a more general approach allowing us to capture more complex patterns among the error terms. We adopt a Multivariate Extreme Value (MEV) model that is derived from the RUM approach. Suppose that the choice set C is partitioned into M overlapping subsets of destinations ($m = 1, .., M$). The CNL model is based on the following probability generating function G:

$$G(e^{\varepsilon_{0n}}, ..., e^{\varepsilon_{Jn}}) = \sum_{m=1}^{M} \left(\sum_{j=0}^{J} (\alpha_{jm}^{\frac{1}{\mu}} e^{\varepsilon_{jn}})^{\mu_m} \right)^{\frac{\mu}{\mu_m}}, \tag{2}$$

with $\alpha_{jm} \geq 0, \frac{\mu}{\mu_m} \leq 1$ and $\forall j$, $\exists m$ such that $\alpha_{jm} \geq 0$.

In this model, the parameters μ_ms capture the similarity between the ε_{jn}s within nest m. The α_{jm} parameters are participation parameters, capturing the extent to which destination j belongs to nest m. In the CNL, μ_m and α_{jm} jointly capture the correlation between the destinations.[7] This specification generalizes the NL approach, in which each destination is assigned to a single nest (i.e., $\alpha_{jm} = 1$ for one m, and 0 for the others). In the CNL specification, this restriction is relaxed. We impose that $\sum_{m=1}^{M} \alpha_{jm} = 1 \ \forall j$. Therefore, the NL model might be seen as a linear restriction of the CNL model. In turn, the logit model can be obtained as a particular case of the NL with $\frac{\mu}{\mu_m} = 1$ for each m.

In addition to its flexibility, the CNL model provides a convenient approach to partition the choice set of destinations into various nests in the case of migration decisions. Each respondent faces a large choice set, comprising more than 200 countries worldwide.[8] The NL model requires this choice set to be partitioned into non overlapping nests. The way this set is initially partitioned is somewhat arbitrary, in terms of the

[7] See Bierlaire, 2006 for a discussion of the conditions to define a GEV function and its properties. In particular this G has properties of non negativity and homogeneity, and complies with some limit properties and the sign of its derivatives. The CDF of the MEV distribution and the expected maximum utility can be directly derived from G.

[8] In fact the number of different ways of partitioning the choice set without priors with J locations in a number of K non-overlapping nests with $1 \leq K \leq (J-1)$ is given by $\sum_{k=1}^{J} \binom{J-1}{K-1}$. In our case study below, the choice set includes 85 foreign countries plus the home location ($J = 86$). The number of partitioning possibilities is equal to $3.68856 \times 10E25$ (including the trivial case of the logit model with a single nest).

number of nests and their composition, making an optimal implementation of the NL approach cumbersome.[9] This is one among the limitations of the NL that have long been emphasized in econometric literature devoted to the modeling of discrete choices (Forinash and Koppelman, 1993). The larger the choice set, the more cumbersome its partitioning in a nested model that relies on non-overlapping nests. A simple way to mitigate the issue consists of reducing the choice set, for example by only considering foreign destinations that have been chosen to a greater extent by respondents (using an arbitrary threshold). Nevertheless, this heuristic method introduces some endogeneity in the estimation that can produce misleading results.[10]

A CNL model that relies on overlapping nests provides an interesting solution to the partitioning of the choice set with a limited number of assumptions. We start by defining broad categories of foreign countries along several dimensions that are usually considered as relevant in migration literature. In our case study below, we define four subgroups of international destinations: OECD versus non OECD countries, English speaking versus non-English speaking, European versus non-European, and contiguous countries versus countries not sharing a common border with India. These categories result in eight overlapping nests of foreign destinations.[11] We then relate each destination to these nests on the basis of objective characteristics through the choice of the α_{jm} parameters. We give the same weight to each nest, which means that for each country, we have four α_{jm}s each equal to $1/4$ and four α_{jm}s each equal to 0. For example, the destination UK has $\alpha_{\text{OECD}} = \alpha_{\text{Eng}} = \alpha_{\text{Eur}} = \alpha_{\text{N.Contig}} = 1/4$ (and $\alpha_{\text{N.OECD}} = \alpha_{\text{N.Eng}} = \alpha_{\text{N.Eur}} = \alpha_{\text{Contig}} = 0$) since it is a European, English-speaking country, and an OECD member state that does not share a border with India. Table 7 in the Appendix provides the values of the α_{jm} parameters for each destination. Interestingly, in contrast to a NL model this structure relies on the same number of nests whatever the size of the choice set.

Clearly, other nesting structures could have been considered. Nevertheless these categories pertain to the main features of the international movements of people. OECD destinations host an overwhelming proportion of immigrants, not only because they

[9] An interesting heuristic approach based on sequential testing of the residuals is proposed in Bertoli and Fernández-Huertas Moraga, 2015. Nevertheless, the total number of potential partitions makes this approach difficult to assess in terms of robustness.

[10] First, the fact that the choice set is exhaustive (i.e., that includes all the alternatives that are considered by individuals) is one of three conditions of discrete choice models (Train, 2009). Furthermore, in discrete choice models such as the ML or the NL, the absolute levels of utility across alternatives are irrelevant and only relative utility matters. Therefore, estimated coefficients reflect *relative* probabilities across alternatives (for example, see Train, 2009, chapter 2). If the least attractive destinations are discarded from the choice set, all coefficients will be affected and the model will be adjusted such that the least visited destination in the selected choice set is the least attractive in the population. This is why, in addition to the 51 destinations for which at least one respondent express a wish to locate, we include in the choice set 34 additional countries that have not been chosen. These 34 countries were selected from the most populous ones. The fact that we do not include all the countries worldwide is only due to computational constraints. For the sake of illustration, with 86 alternatives and more than 30,000 individuals, the average optimization time for the CNL using Biogeme 3.2.6. is over nine hours with a initial set of parameters equal to zero.

[11] Strictly speaking there is therefore a ninth nest including the home destination.

have high income levels, but also because of the quality of their institutions.[12] In general, migration to OECD countries refers to the so-called South-North migration phenomenon, which is of a very different nature to the South-South migration flows between developing countries. The use of English refers to the fact that it is the *lingua franca* and is an important feature for many potential migrants in general, and for Indian migrants in particular. As a former English colony, English is by far the most commonly spoken international language in India. English-speaking countries can also share some common features such as the type of institution or legal systems. The choice of Europe is motivated by the fact that it is a popular destination continent as a whole, and includes a large number of popular destination countries. It also accounts for the fact that when admitted in a European country, immigrants can enjoy more or less free mobility across these destinations, and this can be regarded as an additional source of attractiveness. Last, the contiguity criterion captures the fact that contiguous countries are accessible by different means and are therefore special destinations for some potential migrants who may have friends or relatives on both sides of the border. It also refers to destinations that exhibit some unobserved proximity with India.

Figure 1 illustrates the nested structure of the model used in our case study. The upper part of the model allows us to identify factors that influence the probability of intending to stay in India vs to migrate. The lower part of the model allows us to identify factors of attractiveness across foreign destinations. The circles correspond to the nests of foreign destinations.

Individuals n maximize their utility over the J possible destinations. Under this maximization program, the probability $P_n(j|C)$ that the individual n chooses destination j is given by:

$$P_n(j|C) = Prob(U_{jn} \geq U_{kn} \forall k \in C/\{j\}). \tag{3}$$

This probability can be expressed as:

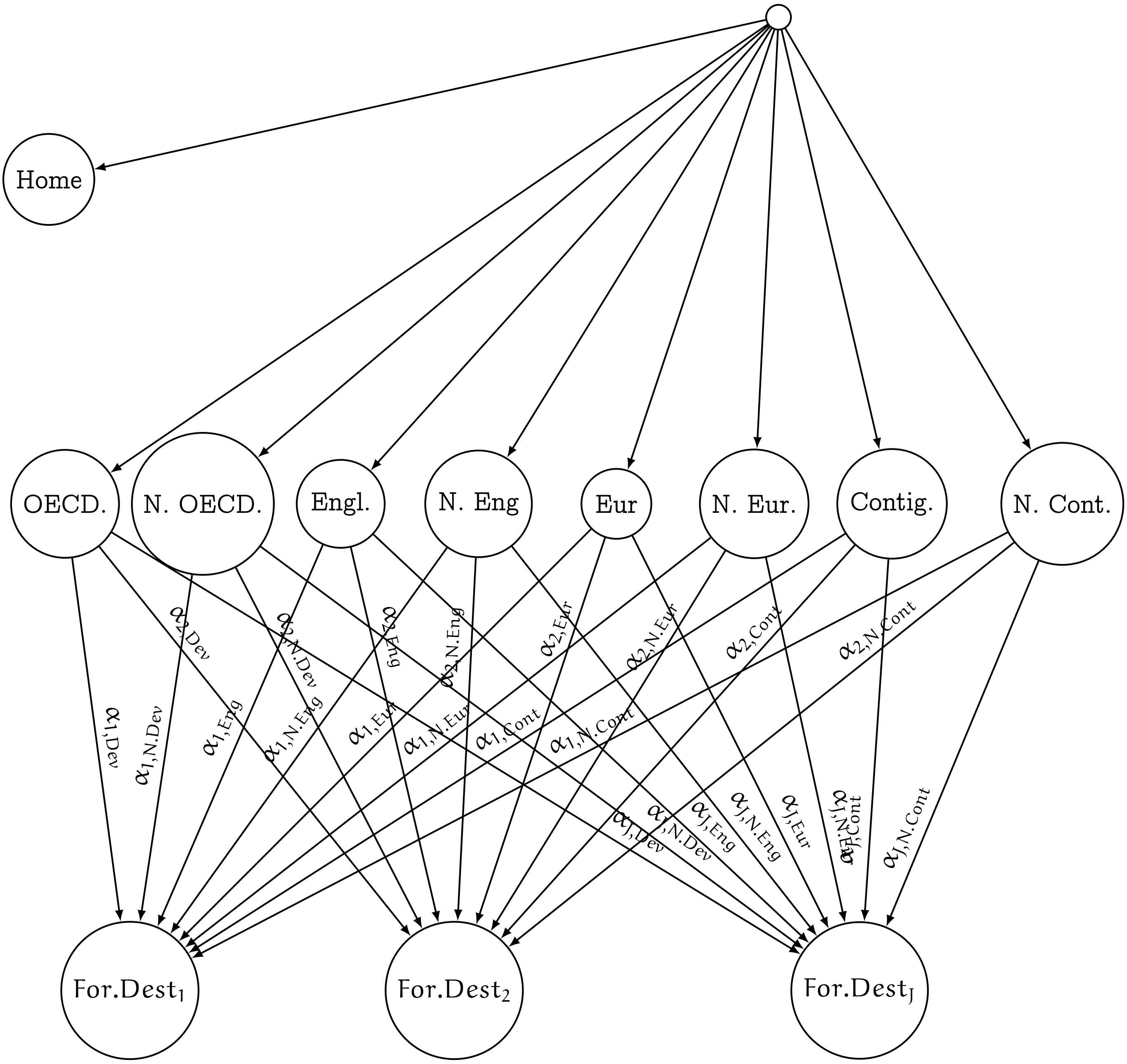
$$P_n(j|C) = \sum_{m=1}^{M} P_n(m|C) P_n(j|m). \tag{4}$$

The probability of choosing a particular destination j can be decomposed into the probability of choosing a particular subset of destinations m and the probability of choosing the destination within the subset m. In turn, using the G function in Eq. (2), the exact form of $P_n(j|C)$ is given by:

$$P_n(j|C) = \sum_{m=1}^{M} \frac{\left(\sum_{j \in C}^{J} \alpha_{jm}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{jn}}\right)^{\frac{\mu}{\mu_m}}}{\sum_{p=1}^{M} \left(\sum_{j \in C}^{J} \alpha_{jn}^{\frac{\mu_n}{\mu}} e^{\mu_n V_{jn}}\right)^{\frac{\mu}{\mu_n}}} \frac{\alpha_{jm}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{jn}}}{\sum_{j \in C}^{J} \alpha_{jm}^{\frac{\mu_m}{\mu}} e^{\mu_m V_{jn}}}. \tag{5}$$

[12] In that regard, since income at destination is included in the deterministic part of the utility, the OECD nest captures the similarity between these destinations in unobserved components such as institutional quality. The type of institutions such as the democratic structures in a country is an explicit condition for joining the OECD "club."

Figure 1: Structure of the CNL model for migration intentions in India.



$$P_n(j|C) = \frac{e^{\mu V_{jn}}}{\sum_{j\in C}^{J} e^{\mu V_{jn}}}. \tag{6}$$

2.2 Deterministic Component of Utility

Without loss of generality, the deterministic part of the total utility in Eq. (1) can be expressed as:

$$V_{jn} = Z'_{jn}\gamma + \delta_{m(j)}, j = 1, ..., J \tag{7}$$

for the utility of moving to foreign country j for individual n and

$$V_{0n} = D'_{n}\beta - \delta_{0}, \tag{8}$$

for the utility of staying for individual n.

The utility of moving depends on a vector Z'_{jn} of destination-specific characteristics X'_{jn} (or X'_{j} if the variable is the same across individuals) interacted with individual characteristics D'_{n}. In our case study, X'_{jn} includes variables reflecting the attractiveness of foreign destinations (such as the level of income per capita, the size of the Indian diaspora and population size) and access to information about them. We also account for dyadic variables such as the geodesic distance between the location of individual n (based on individual's exact location in the origin country) and the destination, and a measurement of religious distance (based on self-reported religious faith and the proportion of the same religious group in the destination country). Some of these variables are interacted with the education level of individual n as a prominent characteristic D'_{n} in order to capture heterogeneous effects across skill groups (Beine et al., 2011, Clemens and Mendola, 2020).

The utility of staying depends on a vector D'_{n} of individual characteristics observed in the home country. These include age, the level of income per household member, the number of children (whether the respondent has at least one child and if so, whether there are more than two), the education level, the type of location at origin (living in a city or not), and the availability of a network abroad (irrespective of its location). The existence of a network is also interacted with the education level to capture heterogeneous sensitivity across skill groups (Beine et al., 2011).

The parameters of interest to be estimated are included in the vectors β and γ. In particular, β is the vector of parameters capturing the influence of individual characteristics on the probability of choosing the home location, as depicted in the upper part of Figure 1, whereas γ is the vector of parameters capturing the influence of destination-specific factors on the probability of choosing that particular foreign destination, as depicted in the middle part of Figure 1. In addition, $\delta_{m(j)}$ is a vector of nest-specific parameters capturing the average attractiveness of the countries belonging to this nest (i.e., a set of nest fixed effects).

2.3 Implied Elasticities and Substitutions

The logit model implies very restrictive substitution patterns. This can be directly seen by computing the change in the probability of choosing a particular location linked to a change in the value of an attribute z_{jn} specific to another location (Train, 2009):

$$\frac{\partial P_n(j|C)}{\partial z_{kn}} = -\gamma_z P_n(j|C)P_n(k|C). \tag{9}$$

The corresponding elasticity is given by:

$$E_{j,z_{kn}} = -\gamma_z z_{kn} P_n(k|C), \tag{10}$$

where γ_z is the estimated effect of covariate z. The cross-elasticity for destination j implied by the logit model is the same across all other destinations (i.e., it does not depend on the specificity of location j). For instance, a given drop in the value of an attribute of destination k for individual n that has a positive impact on the utility ($\gamma_z > 0$) will induce the same proportional increase in the probability of choosing all the other destinations. This pattern of substitution is called *proportionate shifting* and implies that the ratio of the probabilities of two locations stays constant when an attribute specific to a third one changes (for more details, see Train, 2009). It is a manifestation of the IIA property of the logit model at the disaggregated (individual) level.[13] This restriction is lifted in the CNL, implying the ability to assess more complex substitutions across all potential locations. Drawing on Bierlaire, 2006, who studied the theoretical properties of the CNL model, one obtains a corresponding elasticity such as:

$$E_{j,z_{kn}} = z_{kn}[-\gamma_z + \frac{1}{G_j}\frac{\partial G_j}{\partial z_{kn}} - \frac{\partial \ln(\sum_{p \in C} e^{V_p G_p})}{\partial z_{kn}}], \tag{11}$$

where $G_j = \frac{\partial G}{\partial z_{jn}}$ (for more details, see Bierlaire, 2006). Eq. (11) makes it clear that the substitution between destination j and k depends on the characteristics of destination j. For instance, through the G_j terms, it depends on the way the choice set is partitioned, the similarity parameters μ_m and the participation parameters α_{jm}. In other terms, the CNL models allows us to compute substitution rates that are destination specific and that depends on the structure of overlapping nests. Given the analytical complexity of Eq. (11), one needs to compute the elasticities and substitutions at the individual level numerically after estimation.

3 Application: Migration Aspirations in India

We provide a case study that illustrates the relevance of the CNL approach to fit the data and to elicit rich substitution patterns. We use microdata collected in the Gallup World Poll (GWP) surveys on migration aspirations from India over the period 2007-2016. There is a large body of literature investigating the determinants of country-specific or group-specific aspirations to migrate in the fields of demography (Becerra et al., 2010, Becerra, 2012, De Jong et al., 1996, Drinkwater and Ingram, 2009, Wood et al., 2010) and economics (Bertoli and Ruyssen, 2018, Beine et al., 2020, Docquier et al., 2014, Docquier et al., 2019, Dustmann and Okatenko, 2014, Ruyssen and Salomone, 2018, Manchin and Orazbayev, 2014, Manchin and Orazbayev, 2018). The use of stated preferences or contingent valuation surveys to estimate migration aspirations

[13] At the aggregate level, IIA is not satisfied in heterogeneous populations. Therefore, in our estimations of the logit model, we do not have a strict manifestation of the proportionate shifting.

can be criticized (Clemens and Pritchett, 2019). However, it allows us to avoid dealing with attrition and household recomposition issues that are inherent in micro studies (Foster and Rosenzweig, 2002, Bertoli and Murard, 2020). In addition, the recent empirical studies cited above reveal that the stated aspirations are correlated closely with the traditional determinants of migration. Moreover, migration aspirations correlate with actual migration flows (Bertoli and Ruyssen, 2018, Docquier et al., 2019) and amply capture the factors governing individuals' preferences for various destinations. We describe our variables of interest and data sources in Section 3.1, and discuss our empirical findings and implications in Section 3.2.

3.1 Data Sources

This section describes the data used to identify migration aspirations in India, individual characteristics, and destination-specific determinants.

Data on migration aspirations ($P_n(j|C)$). – The GWP database is probably the most comprehensive source of data on migration aspirations worldwide. GWP surveys are conducted in more than 160 countries (representing 99 percent of the world's population aged 15 and over) and are repeated almost every year. Our case study focuses on migration aspirations from India, which is one of the largest countries in the world, and has one of the largest diasporas abroad. According to the United Nations database, the stock of emigrants from India was equal to 12.9 million in 2010 and 15.6 million in 2015. However, these figures represent only 1.6 percent of the Indian population aged 15 and above. In addition, GWP data provide information on migration aspirations from about 3,000 individuals per wave, and on average about 3,500 individuals per year. As the Indian population aged 15 and over is around 975 million, each respondent is representative of about 325,000 individuals. Data is collected by telephone or through face-to-face interviews (Gallup, 2018). The sample of individuals interviewed is designed to be representative of the resident population aged 15 and over. We use data collected between 2007 and 2016. Two waves were conducted during 2009 and during 2012 (we aggregate them per year), while no survey was conducted in 2008. The top panel of Table 1 displays the distribution of respondents over time. Overall, there are more than 35,500 respondents in the nine GWP waves for India, which emphasizes the scale of this survey.

In our analysis, we exploit two specific questions on migration aspirations. The first is: *"Ideally, if you had the opportunity, would you like to move permanently to another country, or would you prefer to continue living in this country?"* For respondents who answered affirmatively, a follow-up question asked about the preferred location: *"To which country would you like to move?"* Combined with data on individual characteristics, the GWP is a rich data source to identify the self-selection factors of international migration and the substitution patterns within the choice set.

Although the data set is cross-sectional and individuals are not followed over time, its annual structure allows us to account for the time variation in migration aspirations. At

Table 1: Distribution of respondents by wave, migration status, and skill level

Year/Group	Number	Percent	Cum.
2007	3,000	8.35	8.35
2009	2,886	8.04	16.39
2010	5,839	16.26	32.65
2011	3,440	9.58	42.23
2012	9,767	27.20	69.43
2013	2,590	7.21	76.65
2014	2,793	7.78	84.43
2015	2,816	7.84	92.27
2016	2,776	7.73	100.00
Total 2007-2016	35,907	100.00	-
Incl. Stayers	33,790	94.10	94.10
Incl. Movers	2,117	5.90	100.00
Incl. Primary (LS)	19,505	54.32	54.32
Incl. Secondary (MS)	12,435	34.63	88.95
Incl. Tertiary (HS)	3,967	11.05	100.00

Note: The stocks/proportions of intended stayers and movers were computed using Gallup question WP1325. The stocks/proportions of respondents by education level were based on Gallup question WP3117. The group "Primary" includes respondents with elementary education or lower (referred to as the low-skilled, LS). The group "Secondary" includes respondents with secondary education completed or up to 3 years of tertiary education (referred to as the medium-skilled, MS). The group "Tertiary" includes respondents with at least 4 years of education completed (referred to as the high-skilled, HS). Individuals refusing to answer question WP1325 and/or failing to give education level are not considered.

the world level, around 20 percent of GWP respondents express a desire to migrate – see Docquier et al., 2014 for an early description of the Gallup data – and this rate is around 40 percent in a few sub-Saharan African countries. In the case of India, as illustrated in the middle panel of Table 1, the mean proportion of aspiring migrants amounts to only 7 percent.[14] In relative terms, India is definitely a case in which mobility, both internally and externally, is low compared with other countries. This has given rise to specific analyses focusing on various explanations such as the existence of insurance networks (Munshi and Rosenzweig, 2016) or of internal borders (Kone et al., 2017). This further stresses the importance of considering stayers in the econometric model, in order to provide a comprehensive analysis of migration intentions.

As documented in previous studies, the set of preferred migration destinations of aspiring migrants differs from the set of actual migrants, and is more concentrated toward high-income OECD countries. India is no exception in this regard, as illustrated in Table 2. We compare the top destinations of aspiring migrants in the left-hand panel

[14] In the estimations of the models, this proportion is further reduced since a subset of intended movers did not indicate their preferred foreign destination in the linked question of the GWP survey.

with those of actual migrants in the right-hand panel. The preferred destinations of aspiring migrants are the U.S. (by far, with 44.3 percent of the total) and the UK (with 9.9 percent of the total). These are followed by the United Arab Emirates (UAE), Singapore, Saudi Arabia, and then other English-speaking OECD countries (Canada and Australia) and Japan. When multiplying the frequency data by the average individual weight (325,000) and dividing them by the number of waves (nine), we obtain an estimate of the total stock of aspiring Indian migrants. This amounts to approximately 52 million people. Turning our attention to actual migrants in 2010, we observe a stock around 13 million (i.e., approximately one quarter of the stock of aspiring migrants), and 21 of the top-30 destinations are identical to those in Col. (1).[15] Countries such as the UAE, the U.S., Saudi Arabia, the UK, Canada and Australia are among the most important destinations. However, the top 15 also includes contiguous countries such as Pakistan, Nepal, and Sri Lanka, as well as additional Persian Gulf countries such as Kuwait, Oman, and Bahrain. In contrast to migration aspirations, the choice of actual destinations is expected to be strongly influenced by out-selection factors such as immigration policies.

Individual characteristics (D_n'). – An appealing feature of the GWP database is that it documents a large set of respondents' personal characteristics, including age, gender, education level, income, family structure, and having a friend or family member abroad (i.e., a personal network link). These characteristics can be used in the modeling of emigration aspirations in Eq. 8. In our empirical analysis, we thus control for individual characteristics that have been described in existing literature as influencing the propensity to emigrate.[16] In particular, we include the log of income per household member in the place of origin (Dao et al., 2018), the existence of a network link abroad (Beine et al., 2011, Munshi, 2004), the family structure (having a child and/or a large family such as more than two children), the age of the respondent (Beine, 2020), and whether the respondent is located in a large city, since international migration occurs primarily from urban areas in developing countries such as India.

We control for respondents' education level to capture the heterogeneity across skill group in the propensity to emigrate. The bottom panel of Table 1 gives the distribution in terms of the three education levels considered in our analysis, referred to as the low-skilled (LS = primary education or lower), the medium-skilled (MS = secondary education completed and up to 3 years of college education), and the high-skilled (HS = at least 4 years of tertiary education completed). With regard to income at origin, we divide the total household income by the equivalent number of household members. We compute this using the *OECD equivalence scale*, which assigns a weight of 1 to the first adult, a weight of 0.7 to other members aged 15 and above, and a weight of 0.5 to members under the age of 15.

These individual characteristics are also useful when modeling the choice of in-

[15] Similar findings emerge when focusing on the year 2015.

[16] Table 9 in Appendix A provides the exact sources of the various individual specific data in the GWP survey data as well as additional descriptive statistics on other characteristics such as employment status.

Table 2: Preferred foreign destinations of intended and actual movers from India

Intended movers 2007-2016 (GWP sample)					Actual emigration stocks in 2010			
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Country	Freq.	People/Year	Percent	Cum.	Country	Freq.	Percent	Cum.
U.S.	640	23,111,111	44.38	44.38	UAE	2,528,605	19.5	19.5
UK	143	5,163,888	9.91	54.29	U.S.	1,990,092	15.35	34.85
UAE	118	4,261,111	8.18	62.47	Saudi Arabia	1,554,650	11.99	46.84
Singapore	82	2,961,111	5.68	68.15	Pakistan	1,395,604	10.76	57.6
Saudi Arabia	79	2,852,777	5.47	73.62	Nepal	698,774	5.39	62.99
Canada	73	2,636,111	5.06	78.68	UK	692,110	5.33	68.32
Australia	69	2,491,666	4.78	83.46	Kuwait	639,379	4.93	73.25
Japan	38	1,372,222	2.63	86.09	Oman	559,558	4.31	77.56
Germany	19	686,111	1.31	87.4	Canada	497,824	3.84	81.4
China	19	686,111	1.31	88.71	Qatar	497,536	3.83	85.23
Nepal	18	650,000	1.24	89.95	Australia	311,116	2.4	87.63
Russia	16	577,777	1.1	91.05	Sri Lanka	307,042	2.36	89.99
Malaysia	15	541,666	1.04	92.09	Bahrain	227,692	1.75	91.74
New Zealand	11	397,222	0.76	92.85	Italy	145,491	1.12	92.86
Switzerland	11	397,222	0.76	93.61	Malaysia	118,008	0.91	93.77
Kuwait	9	325,000	0.62	94.23	Singapore	117,938	0.9	94.67
Bulgaria	9	325,000	0.62	94.85	Germany	65,271	0.5	95.17
Egypt	8	288,888	0.55	95.4	New Zealand	51,020	0.39	95.56
France	7	252,777	0.48	95.88	France	44,677	0.34	95.9
Pakistan	6	216,666	0.41	96.29	Bhutan	43,408	0.33	96.23
Bangladesh	5	180,555	0.34	96.63	South Africa	36,886	0.28	96.51
Lesotho	5	180,555	0.34	96.97	Spain	35,604	0.27	96.78
South Africa	4	144,444	0.27	97.24	Myanmar	34,276	0.26	97.04
Iran	4	144,444	0.27	97.51	Bangladesh	30,881	0.23	97.27
Italy	3	108,333	0.2	97.71	Japan	21,459	0.16	97.43
Cyprus	3	108,333	0.2	97.91	Sweden	19,036	0.14	97.57
Iraq	2	72,222	0.13	98.04	Israel	18,997	0.14	97.71
Qatar	2	72,222	0.13	98.17	Switzerland	18,800	0.14	97.85
Spain	2	72,222	0.13	98.3	Netherlands	17,413	0.13	97.98
Austria	2	72,222	0.13	98.43	Maldives	17,262	0.13	98.11
Hong Kong	2	72,222	0.13	98.56	Hong Kong	16,735	0.12	98.23
Others	18	650,000	1.24	100	Others	209,894	1.61	100
Total	1442	52,072,211	100	-		12,963,038	100	-

Note: In Cols. (1-5), numbers and proportions are computed on usable data from Gallup question WP3120. Unusable answers include refusals, "don't know" answers, and mentioning regions instead of countries or India. The unusable answers amount to 756. In Col. (3), we multiply frequency data by the average individual weight (325,000) and divide it by the number of waves (9). In Cols. (6-9), numbers and proportions are computed on data from the United Nations Population Division.

tended destination (second level of the nested structure) since they are interacted with the destination-specific covariates to identify the parameters of interest. In particular, relevant literature emphasizes the importance of education level as a factor influencing the sensitivity of individuals to most of the important determinants of international migration. This is also the case for income differential (Grogger and Hanson, 2011), distance (Özden et al., 2018), Indian diaspora/network size (Beine et al., 2011, McKenzie and Rapoport, 2011), and religious proximity (Docquier et al., 2019).

Destination-specific variables and interactions (Z'_{jn}). – We supplement and combine the individual characteristics D'_n with destination-specific variables X'_{jn}. These variables capture the deterministic part of the attractiveness of potential foreign destinations in the choice set. These include the main time-varying determinants already identified in existing literature: income level per capita (Grogger and Hanson, 2011), the size of the Indian diasporas (Beine et al., 2011) and population. In contrast to studies on actual migration flows (in which population is used as a proxy for the absorption capacity of the destination country), the effect of population on migration aspirations is more likely to be governed by other factors such as the media coverage and "visibility" of the destination, or an effect of the market size on the variety of goods available to consumers.

These variables can be retrieved from macroeconomic data sources and are observed on an annual basis. We match the year of observation for this data with the year of the GWP wave. For variables that have less frequent observations, such as the Indian network size, we match each GWP wave with observations for the closest year. Income at destination is captured by data for GDP per capita from the Maddison Project database (Bolt et al., 2018), and that are suitable for country comparisons.[17] Data on network size are given by the number of Indian-born individuals living in each destination country, as captured by the estimates of the United Nations data on bilateral migrant stocks (see the right-hand panel of Table 2). Population estimates are retrieved from the United Nations database.

We also include dyadic and time-invariant determinants. We first include a measurement of bilateral distance between the region of residence of individual i and the potential destination. Given that India is a very large country, we use the centroid of the administrative location of each respondent to compute an individual-specific measurement of distance with respect to all potential destinations. For instance, it is well known that people from Western states such as Kerala tend to favor Persian Gulf countries as their foreign location (Clemens et al., 2015). We also account for religious proximity using the declared religion of the respondent in the GWP survey. Religion is one of the main components of culture and cultural proximity has been found to be a factor in attractiveness and selection (Docquier et al., 2019). We create a proximity measurement based on the declared religion of the respondent and the proportions of each broad religion at destination. See Table 10 in Appendix A for more details.

[17] For more details and explanations, see www.ggdc.net/maddison.

3.2 Empirical Results

In this section, we compare the CNL estimation results with those obtained under the NL and logit models. We then discuss the findings of out-of-sample validation experiments. Lastly, we highlight the implications of using a CNL model in terms of elasticities to key variables of interest and substitution patterns.

Estimation results. – Table 3 reports the results of our estimations. For the sake of comparison, we report estimation results of competing models used in relevant literature, as well as estimation results obtained when restricting the sample to aspiring migrants only. Col. (1) provides the results of our CNL for all respondents. Col. (2) provides the results of an NL model in which all foreign destinations are included in the same nest (Buggle et al., 2019, Monras, 2020). The model introduces a natural distinction between stayers and movers, but assumes that IIA holds across all foreign destinations. Col. (3) provides the results of the logit model – the traditional reference used in most of the literature. Col. (4) provides the results obtained when a logit model is estimated on the sample of aspiring migrants (Bertoli and Ruyssen, 2018). In a set of contexts such as ours, it could be argued that ignoring the most popular alternative (i.e., the home location) generates diverging results with respect to the analysis on the full sample. While the Indian context can be seen as an extreme case, with about 95 percent of intended stayers, stayers nevertheless represent the overwhelming majority of respondents in virtually all countries worldwide.[18]

The estimation results of the CNL model,shown in Col. (1) of Table 3, provide new insights into migration aspiration patterns. To start with, the results support the relevance of our nest structure. The CNL generates a strong improvement of the log-likelihood in comparison with the logit and NL models in Cols. (3) and (2). Likelihood-ratio tests suggest that the hypothesis of IIA across foreign alternatives is not supported, as the CNL framework better captures the complexity of substitution patterns across these alternatives. The tests unambiguously reject the relevance of both NL and logit models in favor of the CNL. The estimates of the similarity parameters (the μ_ms) are significantly greater than 1 for most of the nests, confirming the relevance of the choice set partitioning.[19] The estimates of the CNL confirm that there is a higher degree of similarity among OECD destinations, among European destinations, among non-OECD destinations, among non-European destinations, and among non-contiguous destinations.[20] In the NL model of Col. (2), the estimation of the similarity parameter μ_{Foreign} also confirms that foreign destinations are more correlated between each other in

[18] In the GWP data for the year 2017, there are only four countries in which the number of aspiring migrants exceeds the number of intended stayers. The average proportion of aspiring migrants worlwide is slightly greater than 20 percent.

[19] It should be noted that in the CNL and NL models, we normalize $\mu = 1$ so that $\mu_m > 1$ indicates a correlation between alternatives within nest m.

[20] In contrast to the NL models in which there is a one-to-one relationship between the correlation of the error terms and the value of μ, one cannot directly infer the level of correlation from these parameters in the CNL only. This is due to the fact that the patterns of correlation within a nest depend on the participation parameters α_{jm} and the correlation in the overlapping nests.

comparison with the domestic destination. This supports the relevance of a modelling approach such as that adopted by Buggle et al., 2019 and Monras, 2020. One value added aspect of using the CNL is to identify the extent to which the data support the existence of stronger substitution forces within subsets of destinations. This is important from an economic point of view, because it implies that the substitution patterns between foreign destinations depend of their type.[21]

With regard to the estimates of the β parameters governing the utility at origin (V_{0n}), the CNL provides results that are perfectly in line with the existing empirical literature on actual migration flows. We find intuitive results for age (i.e., in line with the standard neoclassical theory, younger people are more willing to emigrate), for the location at origin (i.e., respondents in urban areas are more willing to emigrate), for family structure (i.e., people without children have greater migration aspirations), and for personal network connections (i.e., respondents with a friend or family member abroad are more willing to migrate). In other words, the utility of staying in the home country (V_{0n}) increases with age and with the number of children. By contrast, it decreases with network connections abroad and with the urban nature of the region of origin. An interesting aspect of the estimation is the effect of education level on migration aspirations. In line with Grogger and Hanson, 2011 and McKenzie and Rapoport, 2011, the willingness to move increases with the level of education. Lastly, the effect of income at origin is non-significant, a result that can be driven by the collinearity with education variables but also by the complex nature of the relationship between income and emigration aspirations (Clemens and Mendola, 2020).

Turning our attention to the γ parameters governing the utility associated with a foreign destination (V_{jn}, $j = 1, ..., J$), the results of the CNL show that dyadic migration aspirations increase with the log of income per capita at destination, with the size of the Indian diaspora, and with religious proximity. By contrast, aspirations decrease with geographic distance. These results are in line with the bulk of existing empirical literature. Most of these effects vary with the education level of the respondents. High-skilled respondents are more sensitive to economic conditions (income per capita at destination) and less sensitive to network size, geographic distance, and religious proximity.

One important aspect of our work is to document the implications of using the modeling approach of the stochastic component of the utility. This relates to the findings regarding the impact of observed factors on the probability of emigration and on the intended destination. One of the value added aspects of the CNL is more obvious in the estimation of the δ parameters since both the NL and the logit models fail to account for unobserved correlations across multiple foreign destinations. In addition to the direct effect of income per capita, the CNL allows us to capture the attractiveness of the OECD destinations as reflected by the estimate of the δ_{OECD}, something that is not captured by the NL, or the ML models. Likewise, the CNL captures the relative unattractiveness

[21] the implications of using the gravity models and the issue of multilateral resistance to migration are crucial.

of the non-OECD, non-European, non-English-speaking and non-contiguous countries through the δ_{other}, aspects that the other models also fail to capture.

For illustrative purpose, we also estimate the model on the sample of aspiring migrants only and using the logit model (as in Bertoli and Ruyssen, 2018). It could be argued that disregarding the home location makes the IIA assumption more likely to hold to the extent that foreign destinations are more substitutable between themselves than with the home location. However, the possibility of considering only a subset of alternatives is implied by the IIA being valid between all alternatives (Train, 2009). As a result, this methodology can produce misleading findings, even if the population of interest only includes aspiring migrants. In the case of India, the restriction to the sub-sample of movers implies a huge decrease in the size of the sample as more than 94 percent of the observations are dropped. The results of this estimation are reported in Col.(4) of Table 3. The estimations nevertheless do allow us to gauge how well this model performs empirically, compared with the estimation on the full sample. The results in Col. (4) suggest that the estimations are qualitatively similar to those of the logit model on the full sample. This estimation also fails to capture the attractiveness of OECD countries beyond their impact of observed determinants and the lack of attractiveness of other types of countries.

Table 3: Estimation results across models and samples

	All respondents			Movers only
	(1)	(2)	(3)	(4)
	(CNL)	(NL)	(logit)	(logit)
	Utility of staying in the domestic location ($V_{0n} = D'_n \beta$)			
Age under 65	0.037***	0.037***	0.037***	-
	(0.003)	(0.003)	(0.003)	
Age over 65	-0.045**	-0.044**	-0.045**	-
	(0.022)	(0.022)	(0.022)	
Log of income orig.	-0.060	-0.061	-0.063	-
	(0.041)	(0.041)	(0.041)	
Large city	-0.342***	-0.342***	-0.338***	-
	(0.069)	(0.069)	(0.069)	
No child	-0.193***	-0.195***	-0.195***	-
	(0.063)	(0.063)	(0.063)	
More than 2 children	0.048	0.044	0.057	-
	(0.091)	(0.091)	(0.091)	
Network × LS	-0.891***	-0.900***	-0.879***	-
	(0.135)	(0.135)	(0.136)	
Network × MS	-0.757***	-0.773***	-0.745***	-
	(0.106)	(0.106)	(0.106)	
Network × HS	-0.888***	-0.905***	-0.885***	-
	(0.150)	(0.150)	(0.151)	
Low skilled (LS)	9.12***	6.86***	13.7***	-
	(0.836)	(1.19)	(0.445)	
Medium skilled (MS)	8.78***	6.89***	12.7***	-
	(1.42)	(1.39)	(0.457)	
High skilled (HS)	8.46***	5.99***	12.1***	-
	(0.775)	(1.26)	(0.535)	
	Utility of moving to a foreign location ($V_{jn} = Z'_{jn} \gamma$)			
Log of inc. at dest × LS	0.630***	0.453***	1.090***	1.100***
	(0.113)	(0.136)	(0.108)	(0.107)
Log of inc. at dest × MS	0.711 ***	0.486 ***	1.160***	1.190***
	(0.114)	(0.144)	(0.119)	(0.123)
Log of inc. at dest × HS	0.805***	0.570***	1.360***	1.400***
	(0.151)	(0.179)	(0.186)	(0.186)
Log of diaspora × LS	0.179***	0.098***	0.242***	0.246***
	(0.030)	(0.030)	(0.036)	(0.036)
Log of diaspora × MS	0.190***	0.122***	0.301***	0.305***
	(0.028)	(0.039)	(0.038)	(0.039)
Log of diaspora × HS	0.113***	0.043*	0.108***	0.106**
	(0.031)	(0.023)	(0.046)	(0.046)
Log of distance × LS	-0.684***	-0.529***	-1.310***	-1.280***
	(0.123)	(0.162)	(0.161)	(0.183)
Log of distance × MS	-0.393***	-0.387***	-0.974***	-0.926***
	(0.099)	(0.127)	(0.156)	(0.180)
Log of distance × HS	-0.186	-0.266**	-0.676***	-0.616***
	(0.115)	(0.105)	(0.168)	(0.185)
Religious proximity × LS	0.976***	0.739***	1.400***	1.780***

(Continued on next page)

(Continued from previous page)

	All respondents			Movers only
	(1)	(2)	(3)	(4)
	(CNL)	(NL)	(ML)	(ML)
	(0.142)	(0.211)	(0.175)	(0.262)
Religious proximity × MS	1.130***	0.728***	1.410***	1.670***
	(0.148)	(0.218)	(0.166)	(0.266)
Religious proximity × LS	0.799***	0.610***	0.964***	1.570***
	(0.260)	(0.215)	(0.329)	(0.542)
Log of population	0.339***	0.301***	0.732***	0.732***
	(0.054)	(0.086)	(0.040)	(0.043)
δ_{OECD}	0.382**	-0.078	-0.151	-0.211
	(0.162)	(0.087)	(0.207)	(0.207)
δ_{English}	0.299*	0.823***	2.000***	2.000***
	(0.168)	(0.246)	(0.154)	(0.159)
δ_{European}	-0.081**	-0.082*	-0.219**	-0.193*
	(0.035)	(0.048)	(0.101)	(0.102)
$\delta_{\text{Contiguous}}$	-0.952***	-0.557***	-1.390***	-1.340***
	(0.196)	(0.185)	(0.225)	(0.233)
δ_{Other}	-0.328**	0.032	0.122	0.076
	(0.169)	(0.302)	(0.258)	(0.261)
	Parameters of the nest structure (μ_m)			
μ_{OECD}	2.10***	-	-	-
	(0.287)			
μ_{English}	1.36	-	-	-
	(0.259)			
μ_{European}	11.7***	-	-	-
	(0.298)			
$\mu_{\text{Contiguous}}$	1.92	-	-	-
	(0.731)			
$\mu_{\text{Non-OECD}}$	1.200*	-	-	-
	(0.124)			
$\mu_{\text{Non-English}}$	65.5	-	-	-
	(45.7)			
$\mu_{\text{Non-European}}$	2.28 ***	-	-	-
	(0.532)			
$\mu_{\text{Non-Contiguous}}$	7.34 ***	-	-	-
	(1.93)			
μ_{Foreign}	-	2.45**	-	-
		(0.708)		
Log-Likelihood	-8091.725	-8143.761	-8153.557	-3104.233
LR tests	-	104.08***	123.68***	-
Observations	32492	32492	32492	1295
Parameters	38	31	30	18

Notes: CNL = Cross-nested Logit. NL = Nested logit with nest including all foreign destinations; Logit = Multinomial Logit. Robust standard errors in parentheses: * $p < 0.10$, ** $p < 0.05$,*** $p < 0.01$. For the μ_m parameters, significance levels computed from a one-sided test with null hypothesis: $\mu_m = 1$. LR test gives the test statistics and significance of a Likelihood ratio test comparing against the CNL.

Table 4: Out-of-Sample Validation of the CNL model

	Logit	NL	CNL
Sum Log-Likelihoods	-8187.20	-8177.04	-8148.18

Note: The table reports the sum of the log-likelihoods obtained over the five representative subsamples.

Validation. – One additional way to illustrate the merits of the CNL is to carry out a validation exercise. With that aim, we split the sample randomly into five representative sub-samples of equal size ($\ell = 1, \ldots, 5$). By "representative," we mean that each subsample includes the same proportion of stayers. For each subsample $s = 1, \ldots, 5$, we estimate the three models of columns (1) to (3) of Table 3 on a sample made of all subsamples $\ell \neq s$ and with this model, we calculate the log-likelihood on the subsample $\ell = s$. We then sum up the log-likelihoods obtained for $s = 1, ..., 5$.

Table 4 reports the results of this exercise. Unfortunately, since the summed log-likelihoods do not correspond to the maximum likelihood, it is not possible to conduct standard likelihood ratio tests to discriminate statistically between the models. Nevertheless, the global improvement in the summed likelihoods is important and is of the same order of magnitude as the variations in the likelihoods in Table 3. In addition, the results (not reported here due to considerations of space) show that the CNL leads to an improvement of the likelihood for $s = 1, ..., 5$ (i.e., in all subsamples). The aggregate values reported in Table 4 reflect therefore an improvement of the fit in all out-of-sample portions of the data.

Elasticities. – Most studies of the determinants of international migration are interested in identifying specific determinants and pay particular attention to the values and statistical significance of the parameters. It is therefore important to check whether the estimated marginal effects of the CNL differ from those generated by alternative models. The direct elasticities differ between the logit, NL, and CNL models (Bierlaire, 2006). To illustrate this, we first estimate ARC elasticities and compute the average elasticities of the probability of choosing two destinations, to one of the most important determinants mentioned in existing empirical literature: income per capita at destination. Table 5 reports the estimated elasticities of the number of stayers and of the number of aspiring migrants to the U.S. to a variation of income in the U.S. As the model is non-linear, ARC elasticities are not symmetric and vary with the direction of the income shock. Therefore, we separately report elasticities generated with an increase and with a decrease in income.

The results shown in Table 5 highlight the differences between the elasticities derived from the three competing models. The results in Cols. (1) and (2) show that the logit model clearly overestimates the response in terms of the number of stayers. Compared with the CNL, the NL model implies a lower impact of stayers. Direct elasticities are obtained by looking at the impact on the number of aspiring migrants to the U.S., shown in Cols. (3) and (4) of Table 5. The NL and CNL models that account for

Table 5: Estimated elasticities to a variation in U.S. income

	Δ No. of stayers		Δ No. of asp. migrants to U.S.	
	(1)	(2)	(3)	(4)
	Δ^+ Income	Δ^- Income	Δ^+ Income	Δ^- Income
logit	-0.0190	0.0188	1.16	-1.14
NL	-0.0079	0.0080	0.872	-0.904
CNL	-0.0115	0.0116	0.891	-0.923

Notes: Figures in the table gives the elasticity of the number of aspiring migrants to an increase or decrease in U.S. income per capita computed from the logit, the NL and the CNL model.

deviation from the IIA property yield a lower impact on aspiring migrants compared with the logit. The estimated elasticity is lower by about 25 percent.

Average elasticities are similar between the NL and the CNL models, but average values conceal different patterns at the individual level. In order to visualize the differences, Figure 3 in Appendix C plots pairwise comparisons of the individual ARC elasticities between the models. The individual elasticities are also broken down by level of education, since this dimension is used to generate the heterogeneity of the coefficients in the specification of the deterministic component of utility. The overestimation of the elasticities in the logit model is very obvious in the case of college-educated respondents. The pattern is less obvious for the secondary and primary educated, as a significant proportion of observations lie above the 45 degree line. The NL model tends to yield higher elasticities than the CNL for the college educated, but lower elasticities at lower levels of education. Overall, these differences cancel out in the average values shown in Col. (3) of Table 5.

Substitution patterns. – In addition to improving the quality of fit and the predictive power of the model, the main value added by the CNL technique is that it allows for richer and more complex substitution patterns within the choice set. To illustrate the contribution of the CNL in this regard, we simulate a counterfactual scenario in which the access to the U.S. – the preferred foreign destination for Indian respondents – is removed from the choice set (i.e., the U.S. is no longer accessible to Indian residents).[22] Using estimates from Table 3, we compare how this affects the other alternatives under the CNL, NL and ML models. We first offer a graphical visualization of the substitution patterns. We then discuss the relevance of the CNL findings, and lastly we quantify their implications for migration pressures.

Figure 2 shows the relative changes (left-hand panel) and the absolute changes (right-hand panel) in the number of aspiring migrants for all alternative destinations. The

[22] This counterfactual scenario is far from being a mere academic curiosity. On 27 January 2017, Executive Order 13769 came into effect under President Trump's administration and prevented entries into the U.S. by immigrants from seven Middle East countries (Iran, Iraq, Libya, Somalia, Sudan, Syria and Yemen.

top panel compares the results obtained under the logit model (horizontal axis) with those of the CNL model (vertical axis). By implying IIA across all potential locations, the logit model is expected to spread the former aspiring migrants choosing the US more or less uniformly across alternatives, including the home location. The pattern of substitution is close to a proportional shifting, a clear manifestation of the IIA property (Train, 2009). This is what Figure 2a shows as the logit responses are concentrated around a vertical line.[23] The dispersion measured by the standard deviation of the variation of location choices amounts to 0.58 percent and 0.53 percent, respectively with and without the home location. By contrast, results from the CNL are much more dispersed: the relative variations induced by the CNL range from 1.01 percent (the less substitutable home location) to 70 percent for Canada and 73 percent for Mexico.

The bottom panel compares the NL model (horizontal axis) with the CNL model (vertical axis). While the two models capture the moderate substitution in favor of the home location reasonably well, the patterns predicted by the NL model are similar for the foreign locations. This is what Figure 2c shows, as the NL responses are concentrated around a vertical line, with the exception of the home location. Hence, IIA is more or less satisfied across foreign locations and aspirations to choose non-U.S. alternatives increase almost uniformly by 40 percent, against 1 percent for the home location. The dispersion measured by the standard deviation of the variation of location choices in the NL model amounts to 5.30 percent and 3.26 percent, respectively with and without the home location. For the CNL, the respective values are 22.12 percent and 22.20 percent.

Are the heterogeneous responses predicted by the CNL model meaningful? it should be remembered that the CNL allows each alternative to belong to several (overlapping) nests within which the substitutability between destinations is stronger. The overall substitution forces thus depend on the number of nests shared with the U.S., as well as on the degree of similarity within these nests. Table 6 highlights the differences in substitution forces with five alternatives. 1: Canada, which shares all the nests with the U.S. and is therefore expected to benefit the most of the U.S. borders closing. 2: the UK, which shares three of the four nests with the U.S. 3: Turkey, which also shares three nests with the U.S. but different ones than the UK. 4: Persian Gulf countries, which share only one nest (non-contiguous destinations). 5: The home location, which does not share any common nest with the U.S. In Table 6, Cols. (1) to (3) provide the predictions of the logit, NL, and CNL models, Cols. (4) to (6) provides the results of the counterfactual simulations, and Cols. (7) to (9) quantify the differences.

The top panel of Table 6 focuses on the changes in the number of respondents expressing a desire to emigrate, shown by destination. The logit model seriously underpredicts the number of emigrants choosing Canada. The size of this bias is quite sub-

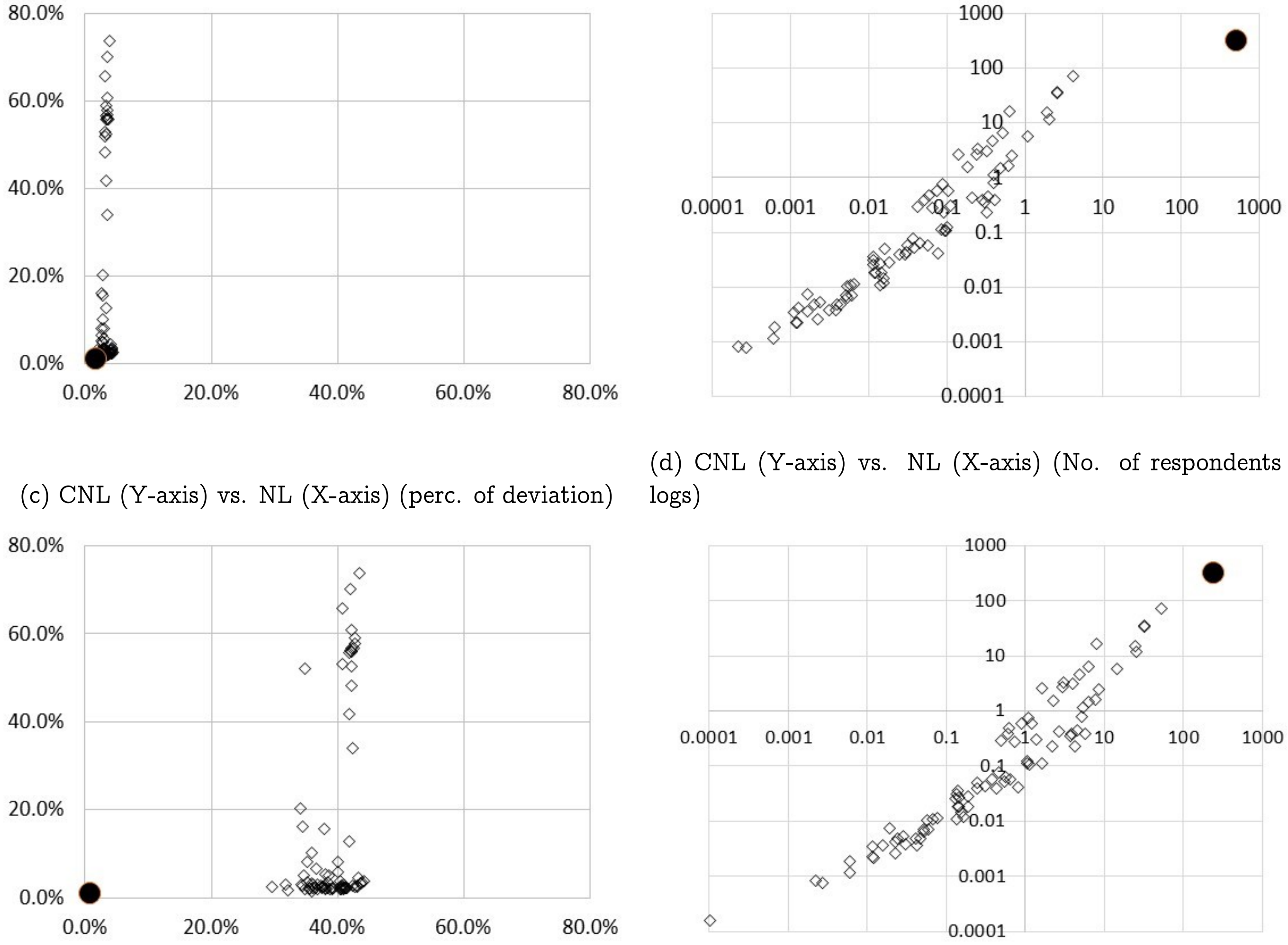
[23] Once again, IIA is not verified at an aggregate level with heterogeneous populations in the logit. The model captures the heterogeneity of the population in terms of age, education, income, personal network link, etc. This explains that when aggregating over alternatives, the proportionate shifting property is not fully satisfied. This is turn implies that the points in the top left-hand panel are not strictly aligned on the vertical axis.

stantial when compared with the prediction of the CNL (59.1 percent versus a 3.5 percent increase with the logit). As expected, by ignoring the specificity of the home location, the logit model over-predicts the number of these potential emigrants who choose to stay in the home country. The NL model provides better results, especially in terms of the predicted variation in the number of stayers. This is understandable given that it accounts for the lower substitution between the home and all the foreign locations. Nevertheless, it ignores the various patterns of substitution across foreign locations and tends to spread the increase uniformly over them. The NL model therefore tends to predict more or less the same variation in intended emigrants between locations that share many common nests with the U.S. and are therefore expected to be close substitutes (such as Canada and Australia) as well as locations that share only one nest (such as Pakistan). It should be emphasized that the similarity across locations is not a monotonic function of the number of common shared nests in the CNL, but we can expect that there is some variation of this type.

Lastly, do these difference matter? In the bottom panel of Table 6, we multiply the predicted stocks and the absolute changes in the number of respondents by their sample weight (i.e., 325,000 divided by nine waves). The numbers are expressed as millions of individuals. When the U.S. becomes inaccessible to 19.7 million aspiring migrants, the CNL model predicts an increase of 11.3 million in the number of stayers, and a residual increase of 8.4 million in the non-U.S. alternatives. By contrast, the logit and NL models predict that the number of stayers increases by 18.5 and 8.8 million, respectively. Hence, the differences between the logit, NL, and CNL models are substantial and predict very different changes in migration pressures. To illustrate, the logit model would imply an increase of intended emigrants choosing the UK of about 150,000 individuals. The NL model would better forecast the total increase, with a predicted increase close to 2 millions. However, it would still underestimate the substitution by about 600,000 individuals as the UK is more similar to the U.S. destination than the average foreign destination. Of course, these variations in the number of aspiring migrants should not be taken as changes in actual migration flows/stocks as there is a huge discrepancy between aspirations expressed in the GWP surveys and actual immigration figures. Additional out-selection factors such as visa restrictions or liquidity constraints further reduce this, to result in the actual number of emigrants to the destination. Although our analysis concerns aspirations, our estimated effects also fit with some basic evidence regarding the impact of H1B visa restrictions on Canadian high-skilled immigration.[24]

[24] For some years,the Canadian immigration authorities have been inviting skilled potential migrants constrained by the quota of the H1B visas to apply for a Canadian visa. The H1B visa quota has been set at 65,000 and has been constantly binding since 2004, prompting an increase in applications from emigrants in STEM occupations such as IT workers or engineers. This has resulted in an increase in entry from many origin countries, including India. Between 2016 and 2017, the inflows in these categories for Indian nationals multiplied approximately threefold, from 11,037 to 36,310. Between 2015 and 2020, the number of Indian Citizens becoming Canadian Permanent Residents jumped from 33,343 to a projected 75,000. These figures also illustrate the belief that many H1B applicants and holders waiting for a green card granting permanent residency on U.S. soil choose Canada as a substitute.

Figure 2: Changes in location preferences under closed U.S. borders, logit, NL and CNL



(a) CNL (Y-axis) vs. logit (X-axis) (perc. of deviation)

(b) CNL (Y-axis) vs. logit (X-axis) (No. of respondents in logs)

(c) CNL (Y-axis) vs. NL (X-axis) (perc. of deviation)

(d) CNL (Y-axis) vs. NL (X-axis) (No. of respondents in logs)

Notes: The panels compare the variations induced by the models under the scenario of a closure of the U.S. borders to Indian respondents. The left-hand panels reproduce the comparison regarding the relative change in preferred locations while the right-hand panels report the change in the number of respondents. The upper panels report the comparison between CNL and NL. The lower panels report the comparison between CNL and ML. The black circle gives the few changes in the preferences for the home location, while the diamonds report the same information for each foreign location.

Table 6: Closing the U.S. border: simulated impact on alternative destinations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
In sample units		Benchmark			Counterfactual			Relative change		
Location	Com. Nests	logit	NL	CNL	logit	NL	CNL	logit	NL	CNL
US	-	535.7	539	545.4	0	0	0	-100%	-100%	-100%
Canada	4	74.9	74.2	61.2	77.5	105.9	97.4	3.47%	42.72%	59.15%
Australia	4	78.61	77.44	60.84	81.21	110.17	95.36	3.31%	42.27%	56.73%
UK	3	127.22	127.29	149.04	131.35	180.98	221.07	3.25%	41.18%	48.32%
Japan	3	20.09	19.88	30.53	20.65	28	46.73	3.21%	40.80%	53.05%
UAE	2	71.61	72.16	94.78	73.54	97.05	110.12	2.69%	34.49%	16.19%
Pakistan	1	14.97	14.31	15.61	15.31	18.87	16.31	2.25%	31.86%	2.88%
Emigrants	-	1294.9	1295	1294.9	782.40	1051.7	980.80	-39.6%	-18.8%	-24.3%
Domestic	0	31197.1	31197	31197.1	31709.6	31440.3	31511.2	1.64%	0.78%	1.01%
In pop units 10E6		Benchmark			Counterfactual			Absolute change		
US	-	19.344	19.463	19.695	0	0	0	-19.344	-19.463	-19.695
Canada	4	2.704	2.679	2.21	2.798	3.824	3.517	0.094	1.145	1.307
Australia	4	2.838	2.796	2.197	2.932	3.978	3.443	0.094	1.182	1.246
UK	3	4.594	4.596	5.382	4.743	6.535	7.983	0.149	1.939	2.601
Japan	3	0.725	0.717	1.102	0.745	1.011	1.687	0.020	0.294	0.585
UAE	2	2.585	2.605	3.422	2.655	3.504	3.976	0.070	0.899	0.554
Pakistan	1	0.54	0.516	0.563	0.552	0.681	0.588	0.012	0.165	0.025
Emigrants	-	46.76	46.763	46.76	28.253	37.978	35.417	-18.507	-8.785	-11.343
Domestic	0	1,126.561	1,126.558	1,126.561	1,145.068	1,135.344	1,137.904	18.507	8.786	11.343

Notes: Figures in Col. (1) report the number of nests shared by the location with the US in the CNL model. Figures in Cols. (2) to (7) give the predicted number of individuals choosing this location under the 2 scenarios. Figures in Cols. (8) to (10) give the predicted change in % in favour of this location with US borders closed (top panel) or absolute changes in the number of aspiring migrants (bottom panel). The bottom of the table express the number in population unit, rather than in sample unit. We multiply frequency data by the average individual weight (325,000) and divide it by the number of waves (9).

4 Conclusion

The question of how people revise their decisions about whether to emigrate, and where to, when facing changes in the global environment is of critical importance in migration literature. For the first time, we propose a cross-nested logit (CNL) approach to generalize the way deviations from the IIA hypothesis can be tested and exploited in migration studies. Substitutability between alternatives is allowed to be stronger within each subgroup of destinations, and each alternative is allowed to belong to several subgroups defined on the basis of consensus dimensions. We provide a case study on migration aspiration data from India, which demonstrates that the CNL approach performs better than standard competing approaches in terms of quality of fit, has stronger predictive power, predicts more heterogeneous responses to shocks, and highlights rich and complex substitution patterns between all possible locations. In particular, we shed light on the low substitutability between the home and foreign destinations,as well as on the subgroups of countries that are considered by potential Indian movers as highly or poorly substitutable.

Although India is a country with a very high proportion of intended stayers by international standards, there is no reason to believe that the results obtained by the estimation of the CNL model are specific to this country. An interesting and appealing feature of the CNL is that the dimensions on which the partitioning of the choice set is made are universal. This means that for any alternative origin country, the same nest structure could be applied to all the other countries. This is a very straightforward advantage over the partitioning of the NL model that would be specific to each country of origin. A second interesting feature of the CNL is that the partitioning strategy is decided ex-ante and does not depend on the number of potential alternatives to be considered. This means, for example, that the approach could be extended to other units of analysis, such as regions. In turn, the CNL approach allows better consideration the pooling across origin countries in order to move the analysis closer to the gravity-like approach that has been used extensively in previous literature. This subsequent step is left for future research.

References

Anderson, J. and van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle, *American Economic Review* **93**(1): 133–160.

Anderson, J. and van Wincoop, E. (2004). Trade costs, *Journal of Economic Literature* **42**(3): 691–751.

Bazzi, S., Burns, S., Hanson, G., Roberts, B. and Whitley, J. (2018). Deterring illegal: Entry: Migrant sanctions and recidivism in border apprehensions, *NBER Working Paper* (25100).

Becerra, D. (2012). The impact of anti-immigration policies and perceived discrimination in the united states on migration intentions among mexican adolescents, *International Migration* **50**(4): 20–32.

Becerra, D., Gurrola, M., Ayon, C. Androff, D., Krysik, J., Gerdes, K., Moya-Salas, L. and Segal, E. (2010). Poverty and other factors affecting migration intentions among adolescents in mexico, *Journal of Poverty* **14**: 1–16.

Beine, M. (2020). Age, Intentions and the Implicit Role of Out-Selection Factors of International Migration., *Working Paper No 8688*, CES Ifo Working Paper.

Beine, M., Bertinelli, L., Cömertpay, R., Litina, A. and Maystadt, J.-F. (2021). The gravity model of forced displacement using mobile phone data, *Journal of Development Economics* (Forthcoming).

Beine, M., Bertoli, S. and Fernández-Huerta Moraga, J. (2016). A practioner's guide to gravity models of international migration, *The World Economy* **39**(4): 496–512.

Beine, M., Docquier, F. and Özden, Ç. (2011). Diasporas, *Journal of Development Economics* **95**(1): 30–41.

Beine, M., Machado, J. and Ruyssen, I. (2020). Do Potential Migrants Internalise Migrant Rights in OECD Host Societies?, *Forthcoming in the Canadian Journal of Economics* .

Ben-Akiva, M. (1973). *The structure of travel demand models*, PhD thesis, Massachusetts Institute of Technology Cambridge, MA.

Ben-Akiva, M. and Bierlaire, M. (1999). Discrete choice methods and their applications to short-term travel decisions, *in* R. Hall (ed.), *Handbook of Transportation Science*, Kluwer, pp. 5–34.

Bertoli, S., Bruecker, H. and Fernández-Huertas Moraga, J. (2016). The European crisis and migration to Germany, *Regional Science and Urban Economics* **60**: 61–72.

Bertoli, S., Bruecker, H. and Fernández-Huertas Moraga, J. (2020). Do processing times affect the distribution of asylum seekers across Europe?, *IZA Discussion Paper* (n. 13018).

Bertoli, S. and Fernández-Huertas Moraga, J. (2013). Multilateral resistance to migration, *Journal of Development Economics* **102**(C): 79–100.

Bertoli, S. and Fernández-Huertas Moraga, J. (2015). The size of the cliff at the border, *Regional Science and Urban Economics* **51**: 1–6.

Bertoli, S., Fernández-Huertas Moraga, J. and Guichard, L. (2020). Rational inattention and migration decisions, *Journal of International Economics* **126**.

Bertoli, S. and Murard, E. (2020). Migration and co-residence choices: Evidence from Mexico, *Journal of Development Economics* **142**: Art. 102330.

Bertoli, S. and Ruyssen, I. (2018). Networks and migrants' intended destination, *Journal of Economic Geography* **18**(4): 705–728.

Bierlaire, M. (2006). A theoretical analysis of the cross-nested logit model, *Annals of Operations research* **144**(1): 287–300.

Bolt, J., Inklaar, R., de Jong, H. and van Zanden, J. L. (2018). Rebasing Maddison: new income comparisons and the shape of long-run economic development, *Technical report*. Maddison Working Paper nr 10.

Bredtmann, J., Klaus, N. and Sebastien., O. (2017). Linguistic distance, networks and migrants' regional location choice, *Technical report*. IZA Discussion paper 11171.

Buggle, J. C., Mayer, T., Sakalli, S. and Thoenig, M. (2019). The refugee's dilemna: Evidence from jewish outmigration in Nazi Germany. Paper presented at the OECD-CEPII Conference on Immigration in OECD Countries.

Cattaneo, C. and Peri, G. (2016). The migration response to increasing temperatures, *Journal of development economics* **122**(C): 127–146.

Chort, I. and Senne, J. (2015). Selection into migration within a household model: Evidence from Senegal, *World Bank Economic Review* **29**: 247–256.

Chort, I. and Senne, J. (2018). You'll be a migrant my son: Accounting for migrant selection within the household, *Economic Development and Cultural Change* **66**(2): 217–263.

Clemens, M. and Mendola, M. (2020). Migration from developing countries: Selection, income elasticity, and Simpson's paradox, *IZA Discussion Papers No 13612*, IZA,Bonn.

Clemens, M., Özden, Ç. and Rapoport, H. (2015). Migration and development research is moving far beyond remittances, *World Development* **65**: 1–5.

Clemens, M. and Pritchett, L. (2019). The new economic case for migration restrictions: An assessment, *Journal of Development Economics* **138**: 153–164.

Dao, T.-H., Docquier, F., Parsons, C. and Peri, G. (2018). Discrete choice models with capacity constraints: an empirical analysis of the housing market of the greater Paris region, *Migration and development: Dissecting the anatomy of the mobility transition* (132): 88–101.

De Jong, G., Richter, K. and Isarabhakdi, P. (1996). Gender, values and intention to move in rural Thailand, *International Migration Review* **30**(3): 748–770.

de la Croix, D., Docquier, F., Fabre, A. and Stelter, R. (2020). The academic market and the rise of universitites in medieval and early modern Europe (1000-1800), *Technical report*, UCLouvain. Manuscript.

Docquier, F., Machado, J. and Sekkat, K. (2015). Efficiency gains from liberalizing labor mobility, *The Scandinavian journal of economics* **117**(2): 303–346.

Docquier, F., Peri, G. and Ruyssen, I. (2014). The cross-country determinants of potential and actual migration, *International Migration Review* **48**: S37–S99.

Docquier, F. and Rapoport, H. (2012). Globalization, brain drain, and development, *Journal of Economic Literature* pp. 681–730.

Docquier, F., Tansel, A. and Turati, R. (2019). Do emigrants self-select along cultural traits? Evidence from the MENA countries, *International Migration Review* **54**(2): 388–422.

Drinkwater, S. and Ingram, P. (2009). How different are the British in their willingness to move? evidence from international social survey data, *Regional Studies* **43**(2): 287–303.

Dustmann, C. and Okatenko, A. (2014). Globalization, brain drain, and development, *Journal of Development Economics* **110**: 52–63.

Dustmann, C., Vasiljeva, K. and Damm, A.-P. (2019). Refugee migration and electoral outcomes, *Review of Economic Studies* **86**(5): 2035–2094.

Forinash, C. and Koppelman, F. (1993). Application and interpretation of the nested logit models of intercity mode choice, *Transportation Research Record Issue* **1413**: 98–106.

Foster, A. and Rosenzweig, M. R. (2002). Household division and rural economic growth, *Review of Economic Studies* **69**: 839–869.

Friebel, G., Manchin, M., Mendola, M. and Prarolo, G. (2019). International migration intentions and illegal costs: Evidence from Africa-to-Europe smuggling routes, *CEPR Dicussion Paper* (n. 13326).

Gallup (2018). Gallup country data set details 2008-2013, available online at http://www.gallup.com/services/177797/country-data-set-details.aspx, *Technical report*.

Gathmann, C. (2007). Effects of enforcement on illegal markets: Evidence from migrant smuggling along the southwestern border, *Journal of Public Economics* **41**(2): 291–315.

Grogger, J. and Hanson, G. H. (2011). Income maximization and the selection and sorting of international migrants, *Journal of Development Economics* **95**(1): 42–57.

Hatton, T. J. (2016). Refugees, asylum seekers, and policy in OECD countries, *American Economic Review* **106**(5): 441–445.

Hatton, T. J. (2017). Refugees and asylum seekers, the crisis in Europe and the future of policy, *Economic Policy* **32**(91): 447–496.

Hatton, T. J. (2020). Asylum migration to the developed world: Persecution, incentives, and policy, *Journal of Economic Perspectives* **34**(1): 75–93.

Jandl, M. (2007). Irregular migration, human smuggling, and the eastern enlargement of the European Union, *International Migration Review* **41**(2): 291–315.

Kone, Z., Liu, M., Matoo, A., Ozden, C. and Sharma, S. (2017). Internal borders and migration in India., *Policy Research Working Paper No 8244*, World Bank Group.

Manchin, M. and Orazbayev, S. (2018). Social networks and the intention to migrate, *World Development* **109**(C): 360–374.

Manchin, M.and Manchin, R. and Orazbayev, S. (2014). Desire to migrate internationally and locally and the importance of satisfaction with amenities, *Technical report*. Paper presented at the FIW-wiiw Seminars in International Economics, April 10, Vienna.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows, *Journal of Population Economics* **23**(4): 1249–1274.

Mayer, T. and Zignago, S. (2011). Notes on CEPII's distance measures: the geodist database, *Technical report*. CEPII Working Paper 2011-25.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behaviour, *in* P. Zarembka (ed.), *Frontiers in Econometrics*, New York: Academic Press, pp. 105–142.

McFadden, D. (1978). Modelling the choice of residential location, *in* A. Karlquist, L. Lundqvist, F. Snickars and J. Weibull (eds), *Spatial Interaction Theory and Residential Location*, North-Hollan, pp. 75–96.

McKenzie, D. and Rapoport, H. (2011). Can migration reduce educational attainment? Evidence from Mexico, *Journal of Population Economics* **24**(4): 1331–1358.

Monras, J. (2020). Economic shocks and internal migration, *CEPR Discussion Papers No DP12977*, CEPR, London.

Munshi, K. (2004). Networks in the Modern Economy: Mexican Migrants in the US Labor Market, *Quarterly Journal of Economics* **118**(2): 549–599.

Munshi, K. and Rosenzweig, M. (2016). Networks and misallocation: Insurance, migration, and the rural-urban wage gap, *American Economic Review* **106**(1): 46–98.

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

Özden, C., Wagner, M. and Packard, M. (2018). Moving for prosperity: Global migration and labor markets., *Policy Research Report Overview No 2018/6/15*, World Bank Group.

Pesaran, M. (2006). Estimation and inference in large heterogeneous panels with a multifactor error structure, *Econometrica* **74**(4): 967–1012.

Ruyssen, I. and Salomone, S. (2018). Female migration: A way out of discrimination?, *Journal of Development Economics* **130**: 224–241.

Train, K. (2009). *Discrete Choice Methods with Simulation*, Cambridge University Press.

Vovsha, P. (1997). Application of cross-nested logit model to mode choice in Tel-Aviv, Israel, metropolitan area, *Transportation Research Record* **1607**: 6–15.

Wood, C., Gibson, C., Ribeiro, L. and Hamsho-Diaz, P. (2010). Crime victimization in Latin America and intentions to migrate to the United States, *International Migration Review* **44**(1): 3–24.

Appendices

Appendix A Participation parameters $\alpha_{j,m}$

Table 7: Values of participation parameters $\alpha_{j,m}$

Destination	$\alpha_{j,\mathrm{OECD}}$	$\alpha_{j,\mathrm{Eng}}$	$\alpha_{j,\mathrm{Eur}}$	$\alpha_{j,\mathrm{Contig}}$	$\alpha_{j,\mathrm{N.OECD}}$	$\alpha_{j,\mathrm{N.Eng}}$	$\alpha_{j,\mathrm{N.Eur}}$	$\alpha_{j,\mathrm{N.Cont}}$
United States	0.25	0.25	0	0	0	0	0.25	0.25
Egypt	0	0	0	0	0.25	0.25	0.25	0.25
Morocco	0	0	0	0	0.25	0.25	0.25	0.25
Saudi Arabia	0	0	0	0	0.25	0.25	0.25	0.25
Jordan	0	0	0	0	0.25	0.25	0.25	0.25
Pakistan	0	0	0	0.25	0.25	0.25	0.25	0
Indonesia	0	0	0	0	0.25	0.25	0.25	0.25
Bangladesh	0	0	0	0.25	0.25	0.25	0.25	0
United Kingdom	0.25	0.25	0.25	0	0	0	0	0.25
France	0.25	0	0.25	0	0	0.25	0	0.25
Germany	0.25	0	0.25	0	0	0.25	0	0.25
Spain	0.25	0	0.25	0	0	0.25	0	0.25
Italy	0.25	0	0.25	0	0	0.25	0	0.25
Sweden	0.25	0	0.25	0	0	0.25	0	0.25
Denmark	0.25	0	0.25	0	0	0.25	0	0.25
Iran	0	0	0	0	0.25	0.25	0.25	0.25
Hong Kong	0	0	0	0	0.25	0.25	0.25	0.25
Singapore	0	0	0	0	0.25	0.25	0.25	0.25
Japan	0.25	0	0	0	0	0.25	0.25	0.25
China	0	0	0	0.25	0.25	0.25	0.25	0
Kenya	0	0	0	0	0.25	0.25	0.25	0.25
South Africa	0	0.25	0	0	0.25	0.25	0.25	0
Canada	0.25	0.25	0	0	0	0.25	0.25	0
Australia	0.25	0.25	0	0	0	0.25	0.25	0
Sri Lanka	0	0	0	0.25	0.25	0.25	0.25	0
Thailand	0	0	0	0	0.25	0.25	0.25	0.25
New Zealand	0.25	0.25	0	0	0	0.25	0.25	0
South Korea	0.25	0	0	0	0	0.25	0.25	0.25
Russia	0	0	0.25	0	0.25	0.25	0	0.25
Austria	0.25	0	0.25	0	0	0.25	0	0.25
Brunei	0	0	0	0	0.25	0.25	0.25	0.25
Bulgaria	0.	0	0.25	0	0.25	0.25	0	0.25
Cyprus	0	0	0.25	0	0.25	0.25	0	0.25
Djibouti	0	0	0	0	0.25	0.25	0.25	0.25
Finland	0.25	0	0.25	0	0	0.25	0	0.25
Gabon	0	0	0	0	0.25	0.25	0.25	0.25
Iraq	0	0	0	0	0.25	0.25	0.25	0.25
Kiribati	0	0	0	0	0.25	0.25	0.25	0.25
Kuwait	0	0	0	0	0.25	0.25	0.25	0.25
Lesotho	0	0	0	0	0.25	0.25	0.25	0.25
Libya	0	0	0	0	0.25	0.25	0.25	0.25
Malaysia	0	0	0	0	0.25	0.25	0.25	0.25

Continued on next page

Continued from previous page

Destination	$\alpha_{j,\mathrm{Devel}}$	$\alpha_{j,\mathrm{Eng}}$	$\alpha_{j,\mathrm{Eur}}$	$\alpha_{j,\mathrm{Contig}}$	$\alpha_{j,\mathrm{N.Devel}}$	$\alpha_{j,\mathrm{N.Eng}}$	$\alpha_{j,\mathrm{N.Eur}}$	$\alpha_{j,\mathrm{N.Cont}}$
Maldives	0	0	0	0	0.25	0.25	0.25	0.25
Mauritius	0	0	0	0	0.25	0.25	0.25	0.25
Mongolia	0	0	0	0	0.25	0.25	0.25	0.25
Nepal	0	0	0	0.25	0.25	0.25	0.25	0
Qatar	0	0	0	0	0.25	0.25	0.25	0.25
Swaziland	0	0	0	0	0.25	0.25	0.25	0.25
Switzerland	0.25	0	0.25	0	0	0.25	0	0.25
Tunisia	0	0	0	0	0.25	0.25	0.25	0.25
UAE	0	0	0	0	0.25	0.25	0.25	0.25
Bhutan	0	0	0	0.25	0.25	0.25	0.25	0
Myanmar	0	0	0	0.25	0.25	0.25	0.25	0
Afghanistan	0	0	0	0	0.25	0.25	0.25	0.25
Angola	0	0	0	0	0.25	0.25	0.25	0.25
Argentina	0	0	0	0	0.25	0.25	0.25	0.25
Brazil	0	0	0	0	0.25	0.25	0.25	0.25
Ivory Coast	0	0	0	0	0.25	0.25	0.25	0.25
Cameroon	0	0	0	0	0.25	0.25	0.25	0.25
DR Congo	0	0	0	0	0.25	0.25	0.25	0.25
Columbia	0	0	0	0	0.25	0.25	0.25	0.25
Algeria	0	0	0	0	0.25	0.25	0.25	0.25
Ethiopia	0	0	0	0	0.25	0.25	0.25	0.25
Gabon	0	0	0	0	0.25	0.25	0.25	0.25
Madagascar	0	0	0	0	0.25	0.25	0.25	0.25
Mozambique	0	0	0	0	0.25	0.25	0.25	0.25
Niger	0	0	0	0	0.25	0.25	0.25	0.25
Nigeria	0	0	0	0	0.25	0.25	0.25	0.25
Peru	0	0	0	0	0.25	0.25	0.25	0.25
Philippines	0	0	0	0	0.25	0.25	0.25	0.25
Poland	0.25	0	0.25	0	0	0.25	0	0.25
Romania	0	0	0.25	0	0.25	0.25	0	0.25
Sudan	0	0	0	0	0.25	0.25	0.25	0.25
Syria	0	0	0	0	0.25	0.25	0.25	0.25
Tanzania	0	0	0.25	0	0.25	0	0.25	0.25
Uganda	0	0	0.25	0	0.25	0	0.25	0.25
Ukraine	0	0	0.25	0	0.25	0.25	0	0.25
Uzbekistan	0	0	0	0	0.25	0.25	0.25	0.25
Venezuela	0	0	0	0	0.25	0.25	0.25	0.25
Vietnam	0	0	0	0.25	0.25	0.25	0.25	0
Yemen	0	0	0	0	0.25	0.25	0.25	0.25
Mexico	0.25	0	0	0	0	0.25	0.25	0.25
Turkey	0.25	0	0.25	0	0	0.25	0	0.25
Bahrein	0	0	0	0	0.25	0.25	0.25	0.25
Oman	0	0	0	0	0.25	0.25	0.25	0.25

Appendix B Data: additional information

B.1 Employment status

Table 8: Distribution of employment status

Employment Status	Freq.	Percent	Cum.
Employed full time for an employer	8,653	26.60	26.60
Employed full time for self	5,634	17.32	43.92
Employed part time/Do not want full time	1,180	3.63	47.54
Unemployed	1,425	4.38	51.92
Employed part time/Want full time	1,174	3.61	55.53
Out of the workforce	14,466	44.47	100.00
Total	32,532	100.00	-

B.2 Individual-specific variables from the GWP survey (D'_n)

Table 9 gives the sources from the GWP survey and describes the main individual-specific variables.

B.3 Sources and explanation of destination-specific variables (X'_{jn}).

Table 10 gives the sources and describes the main country-specific variables.

Table 9: Sources and description of individual-specific variables in the GWP survey

Variable	Description	Gallup question code
Age	Age of the respondent	WP1220
Education level	Education level of the respondent	WP3117
Large city	Dummy= 1 if respondent lives in a large city	WP14
Location Suburb	Dummy= 1 if respondent lives in suburb of a large city	WP14
Network abroad	Dummy= 1 if respondent has a network abroad	WP3333
Income per capita	Average income in the household	INCOME4
Number of children	Number of children under 15 in the household	WP1230
number of other adults in household	Number of members over 15 in household	WP12
Income	Computed income of individual	Authors calculation based on equivalence scale ♣
Location	Indian state of location of individual	RegionIND (for waves 2008-2016)
Location	Large Region of location of individual	Region2IND (for wave 2007)
Religion	Religion of individual	Aggregation based on WP1233 ♣

♣: see specific subsections in this Appendix for further explanations.

Table 10: Sources and description of destination-specific variables

Variables	Description	Source	Time Frequency
Income	GDP per capita	Maddison Database v. 2018.	Annual
Diaspora	Total Stock of Indian-born individuals	United Nations - International Migrant Stock in 2019	5-year
Population	Total population at destination	United Nations Population prospects in 2019	5-year
Distance	Distance between India and destination	Geodist CEPII database Mayer and Zignago, 2011	-
Visa	Visa restriction between India and destination	DEMIG-ATA restriction	Annual

B.4 Computation of individual income

Individual income in the origin country is inferred from the household's income using an income equivalence scale. We first retrieve the household income per capita hc_n (Gallup question INCOME4). We obtain the total income of household (h_n) by multiplying income per capita by the number of household members (HHSIZE variable in Gallup). We then use information about household composition in terms of the number of other adults living in the households $adults_i$ (defined as other members over the age of 15 and given by the variable WP12 in Gallup) and the number of children (defined as the number of other members under the age of 15, given by Gallup variable WP1230). Individual income I_n is given by $I_n = \frac{h_n}{1+0.5adults_n+0.3children_n}$.

B.5 Computation of individual geodesic distances

Individual distance from the foreign destination j d_{jn} is based on the location of individual n in a given Indian state. It is a computed distance between centroids of $state_n$ and $country_j$. For data collected over the 2008-2016 period, the location of individual n is retrieved from the Gallup variable REGIONIND, and indicates which of the 25 Indian States an individual originates from. We use the geodesic coordinates of the capital of each state and compute the distance between this capital and the capital of each country j based on the great circle distance using the *geodist* command in Stata. For the 2007 wave of Gallup data, since the location in each state is unavailable, location is expressed in terms of larger regions (five regions: East, West, North, South and Central Region captured by REGION2IND in Gallup). We follow the same procedure but ascribe a centroid to each of the region.[25]

B.6 Computation of individual religious proximity

We also compute a measurement of religious proximity for each individual with respect to each country j. This measurement is the computed probability of an individual n of religion r randomly meeting a resident of the same religion in destination j. We first retrieve the reported religion of individual n using Gallup question WP1233 . We aggregate the 25 religions reported in Gallup into four large categories. Table11 reports the aggregation process for the main reported religions.

[25] The corresponding capitals are: Bophal for Central Region, Chennai for the South, Dehli for the North, Calcutta for the East, and Pune for the West.

Table 11: Aggregation of reported religions.

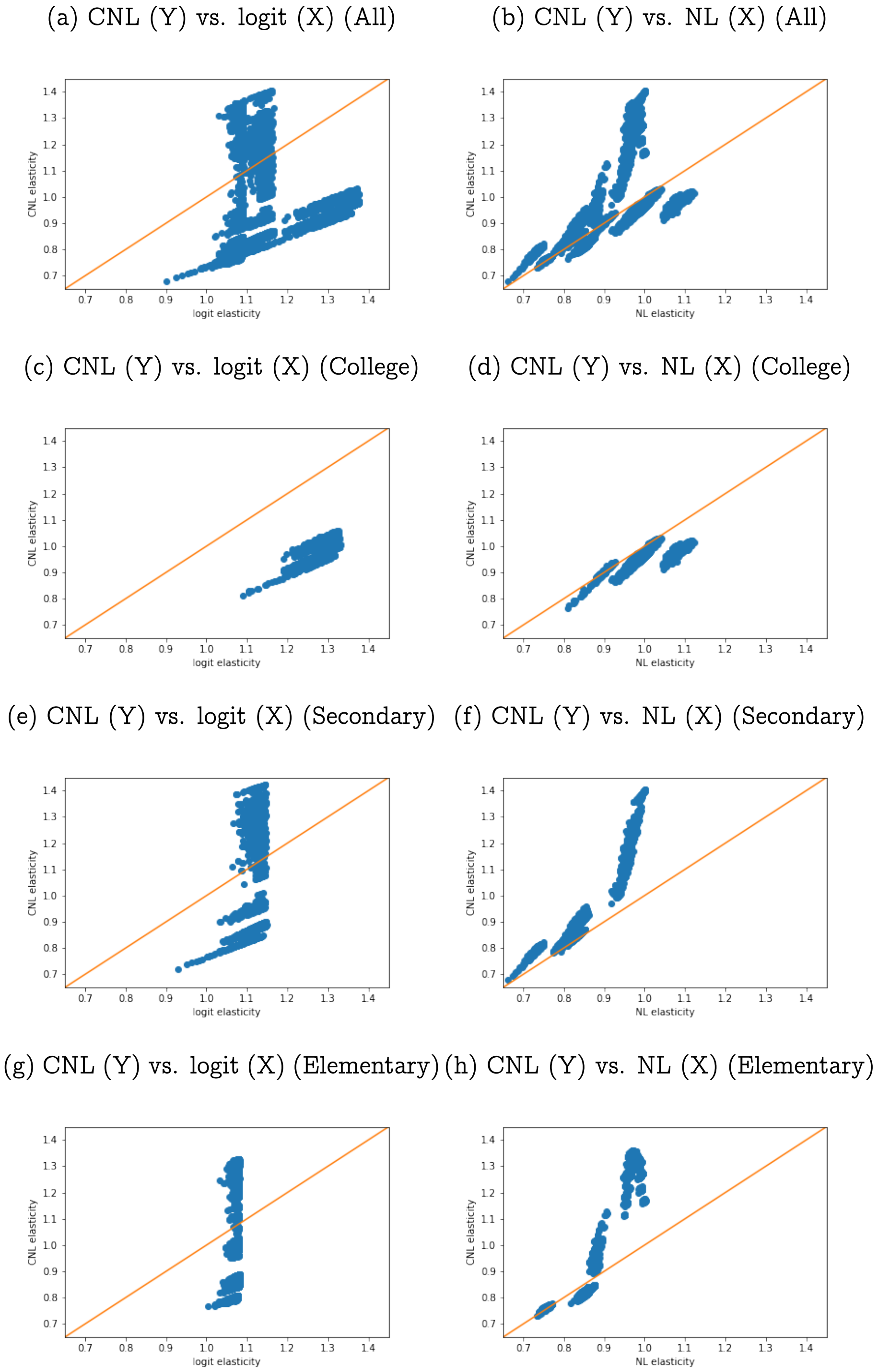
Gallup reported religion	Aggregate religion	Proportion (as %)
Roman Catholic	Christian	0.95
Protestant/Anglican	Christian	0.39
Orthodox	Christian	0.32
Christian	Christian	0.35
Islam	Muslim	6.65
Shiite	Muslim	0.76
Sunnite	Muslim	4.72
Druze	Muslim	0.02
Hinduism	Hinduism	82.67
Other	Other	0.06
Buddhism	Other	0.67
Sikhism	Other	2.14
Jainism	Other	0.13
Other reported	Other	0.16

Other reported religions include African Traditional, Confucianism, Spiritism, Shinto, Zoroastrianism, Rastafarianism, Others, refusals to answer, and unknown answers.

The proportion of religious practitioners of these four large religion groupings in each country is retrieved from the emphInfoplease.com website. The information is completed by the information provided on each country's Wikipedia page if the starting information is not precise enough. This proportion of religious practitioners of the reported religion of individual n provides the measure of religious proximity.

Appendix C Individual ARC elasticities to US income

Figure 3: Comparison of individual ARC elasticities to US income across models



(a) CNL (Y) vs. logit (X) (All)

(b) CNL (Y) vs. NL (X) (All)

(c) CNL (Y) vs. logit (X) (College)

(d) CNL (Y) vs. NL (X) (College)

(e) CNL (Y) vs. logit (X) (Secondary)

(f) CNL (Y) vs. NL (X) (Secondary)

(g) CNL (Y) vs. logit (X) (Elementary)

(h) CNL (Y) vs. NL (X) (Elementary)

Note. Each figure plots the joint values of individuals ARC elasticities implied by the estimates of equation (3) for two models. ARC elasticities capture the change in the number of intended migrants to the US to a 10% increase in U.S. income. The first column gives the elasticities of the CNL versus the logit. The second column gives the elasticities of the CNL versus the NL. Rows vary by education level of the respondents.Rows (1), (2), (3) and (4) report joint values for respectively all respondents, college educated, respondents with secondary education and those with primary education.