Sampling of alternatives in migration aspiration models

Michel Beine

Michel Bierlaire

Evangelos Paschalidis

Andreas B. Vortisch

STRC conference paper 2023 **April 14, 2023**

STRC | **23rd Swiss Transport Research Conference**
Monte Verità / Ascona, May 10-12, 2023

Sampling of alternatives in migration aspiration models

Michel Beine
Department of Economics and Management
University of Luxembourg
michel.beine@uni.lu

Michel Bierlaire
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne, Switzerland
michel.bierlaire@epfl.ch

Evangelos Paschalidis
Transport and Mobility Laboratory
École Polytechnique Fédérale de Lausanne, Switzerland
e@mail.web

Andreas B. Vortisch
Department of Economics and Management
University of Luxembourg
andreas.vortisch@uni.lu

April 14, 2023

Abstract

The use of discrete choice models (DCMs) is a regular approach to investigating migration aspirations concerning destination choices. However, given the complex substitution patterns between destinations, more advanced model specifications than the multinomial logit (MNL) and nested logit (NL) models which are commonly found in the literature are required. The cross-nested logit (CNL) model allows for a more sophisticated representation of the stochastic structure of destination choices, through the use of overlapping nests while it is also addressing deviations from the property of independence of irrelevant alternatives. However, the shift towards CNL does not come without a cost; these models can be computationally expensive to estimate, especially as the number of observations increases. The estimation speed can be mitigated though via sampling of alternatives i.e. reducing the number of alternatives in the model specification. This method has been previously used mostly in the context of residential choice location. In the current work, we implement sampling of alternatives on migration aspiration choices using the Gallup World Poll data. We examine the impact of stratification and number of alternatives on the CNL model estimates. Moreover, we consider additional MNL and NL specifications to further understand the implications of sampling on DCMs used for modelling migration aspirations.

Keywords

Migration, aspiration, logit, sampling of alternatives

Contents

List of Tables

1 Introduction

International migration is a crucial but controversial topic that raises issues and concerns to be resolved in the parties involved. In industrialised nations, the proportion of immigrants in total population increased from 4.5 to 12 percent between 1960 and 2019, raising fears about economic costs for natives, loss of national identity, and integration issues. In poor countries, international migration raises concerns regarding brain drain of highly-skilled people, as college and university graduates have a higher tendency to emigrate than the less educated (Beine *et al.*, 2021). The accurate understanding and prediction of migration motifs are crucial from a policy-making perspective, to measure for instance the immigration pressure a country faces. Correct estimates of migration flows are very relevant in situations of drastic policies a country implements and their impact on third countries. This analysis is highly relevant for policymakers around the globe because political viewpoints on immigration policies may be shifting and induce more restrictive policies abroad with domestic repercussions. Therefore, it is imperative to better understand migration aspirations, how they translate into actual migration numbers, and the impact of immigration policies on third countries. Improved models can help to predict and adjust migration flows but also to identify and mitigate adverse effects on all parties involved.

The questions of how many people migrate, which people are more likely to migrate, and where migrants choose to move have been extensively investigated in recent literature. However, many of the previous studies that focused on migration aspiration and used discrete choice models (Bekaert *et al.*, 2021; Docquier *et al.*, 2020; Bertoli and Ruyssen, 2018; Gubert and Senne, 2016; Lovo, 2014 to name a few) omitted correlation across destinations. This specification assumes independence from irrelevant alternatives (IIA) which implies that cross elasticities due to the change of an attribute in one destination are identical for all alternatives. However, individuals are expected to substitute their choice in favour of certain locations instead of other locations. This behaviour could be for instance due to factors such as language, religion, visa restrictions and others. Other studies presented more elaborated model specification to address the deviations from the IIA. The most popular approach to capture substitution patterns between countries is the nested logit model (as in Bertoli and Moraga, 2013; Ortega and Peri, 2013; Buggle *et al.*, 2020; Monras, 2018; Langella and Manning, 2021). Beine *et al.* (2021), implemented a cross-nested logit (CNL) model to approximate migration aspirations of Indian individuals.

Models with more sophisticated functional forms and prediction accuracy can significantly

assist to forecast and adjust migration flows on such occasions. However, complex model structures such as the CNL pose estimation challenges as they can be computationally expensive and require a considerable amount of time. This issue is magnified in the presence of a large set of alternatives which is also the case in the area of migration; potential emigrates have a choice set of all countries globally (approximately 230). The problem of large choice sets is not new in the field of discrete choice modelling. To tackle this issue, McFadden (1977) developed the method of sampling of alternatives demonstrating that with the addition of a correction term in the utility function it was feasible to obtain consistent parameter estimates. Bierlaire *et al.* (2008) extended the initial sampling of alternatives framework for generalised multivariate extreme value (MEV) models which was later implemented by Guevara and Ben-Akiva (2013). In literature, there are examples of studies attempting to investigate the impact of alternative sample size in the implementation of sampling of alternatives (Lai and Bierlaire, 2015; Tsoleridis *et al.*, 2022; Nerella and Bhat, 2004). However, to the best of our knowledge the concept of sampling of alternatives and the implications of the sample size of alternatives has been never implemented and examined in the concept of migration. In the present paper, we are attempting to cover this gap in the literature. In particular, we are building on the approach and migration aspiration models presented by Beine *et al.* (2021). Following the original model specifications of the aforementioned work, we are investigating the impact of sampling of alternatives on the parameter estimates of a series of discrete choice models.

The remainder of the paper is organised as follows; Section 2 highlights the data that we used. Section 3 percents our methodological framework and sampling of alternatives approach. Section 4 focuses on the current findings. The paper concludes in Section 5 with some relevant discussion, limitations and next steps.

2 Data

In the current paper we are building on the case study presented by Beine *et al.* (2021) who developed a CNL model for migration aspiration in India. The authors estimated their models using the Gallup World Poll (GWP) surveys on migration aspirations from India over the period 2007-2016. The GWP database is probably the most comprehensive cross-sectional 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. Migration aspiration is captured via the question "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?". Respondents replying positively were asked a follow-up question "To which country would you like to move?".

A very interesting finding was that the only 7% of respondents aspired to migrate which is consistent with migration mobility in India. Hence, *stayers* were also included in the original model specification however, with a different utility function compared to the *movers*. From the latter, 44.3% reported USA as their preferred destination followed by UK (9.9%). These were followed by the United Arab Emirates (UAE), Singapore, Saudi Arabia, and then other English-speaking OECD countries (Canada and Australia) and Japan.In total, 51 different countries were reported as potential destinations. To these, Beine *et al.* (2021) added 34 additional countries in their model specification.

On top of migration aspiration, the GWP data also include information regarding individual characteristics. The most interesting that were also used in the original work of Beine *et al.* (2021) are: Level of education/skill where 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) were defined. Other notable variables are income (log of income per household member used in the model), existence of a network link abroad, family structure (number of children, if any), age, gender and whether the respondent was located in a large city.

The individual characteristics from the GWP data were further augmented with interactions with destination-specific variables. These variables capture the deterministic part of the attractiveness of foreign destinations in the choice set. Some of these variables were time varying (the year each respondent took the survey was used as reference) such as income per capita and size of Indian diasporas in each destination. Also the distance of each respondent's region to the capital of each potential destination was calculated.

After all data processing and cleaning, 32,942 observations were considered for model estimation.

3 Methodology

3.1 Modelling framework

Under the assumption that migration aspiration choices are based on utility maximisation, we can then define utility in the form of a linear function as:

$$U_{in} = V_{in} + \epsilon_{in}, \tag{1}$$

where V_{in} represents the deterministic part related to an individual n for a potential location i, while ϵ is and i.i.d. Gumbel distributed disturbance term.

The utility function in Equation 1 which is typically used in the specification of the MNL can be also used in the probability formulation of Multivariate Extreme Value (MEV) models, a family of models based on the use of extreme value distributions. The probability specification of MEV is

$$P_{in} = \frac{e^{V_{in}+lnG_j(e_j^V)}}{\sum_j e^{V_j n+lnG_j(e_j^V)}} \tag{2}$$

In order to capture more complex patterns among the disturbance terms we adopt a Multivariate Extreme Value model (MEV) that stems from the random utility approach. Under the assumption that a choice set C is divided into M overlapping subsets of destination choices ($m = 1, \ldots, M$), we can derive the CNL model specification as

$$G_j(e_j^V) = G(e^{V_{0n}}, ..., e^{V_{In}}) = \sum_{m=1}^{M} \left(\sum_{j=0}^{J} (\alpha_{jm}^{\frac{1}{\mu}} e^{V_{jn}})^{\mu_m} \right)^{\frac{\mu}{\mu_m}}, \tag{3}$$

where $\alpha_{jm} \geq 0$, $\frac{\mu}{\mu_m} \leq 1$ and $\forall$j, $\exists$m such that $\alpha_{jm} \geq 0$. In the specification of Equation 3, α_{jm} is a participation parameter and μ_m captures the similarity of unobserved attributes across countries in nest m. The α_{jm} are participation parameters denoting the extent to which a destination j is part of nest m. In the CNL model, μ_m and α_{jm} capture the correlation between destinations. The nested logit (NL) model is a special case of Eq. (3) if a parameter α_{jm} takes a non-zero value for a nest while it is zero for the other nests. Moreover, in case of an NL, where $\frac{\mu}{\mu_m} = 1$, the specification collapses to a multinomial logit model (MNL).

Two different specifications for the utility function were considered depending on whether an individual was aspiring to migrate or not. In case of aspiration, the utility specification included destination-related variables and interactions with respondents' characteristics. On the other hand, the utility specification for stayers included socio-demographic variables such as skill-level, age, number of children and size of residence location.

3.2 Sampling of alternatives

The idea behind the sampling of alternatives is the reduction of alternatives used in the model specification to reduce the computational cost while at the same time. However, a naive selection of alternatives can be considered a case of model misspecification leading to biased estimates (Swait and Ben-Akiva, 1987). As stated earlier in the introduction section, McFadden (1977) addressed this issue via the introduction of a sampling correction (SC) term and showed that this approach can lead to unbiased estimates (bias defined as the difference between the sampled estimates and the estimates obtained using the full choice set).

For a typical multinomial logit model, adding the correction term modifies the probability as

$$P_{in}|D_n = \frac{e^{V_{in}+ln\pi(D_n)}}{\sum_{j\in D_n} e^{V_{jn}+ln\pi(D_n)}} \tag{4}$$

The additional term is is the logarithm of the probability of creating the choice set given that alternative i was chosen for individual n. That can be also considered as a penalty added to the utility, since $\pi(D_n)$ is constrained between 0 and 1 hence its logarithm will always be a negative value (smaller probability of sampling a choice set results in bigger penalty in the utility function). Following the MEV choice probability presented in Equation 2, Lai and Bierlaire (2015) simplified the sampling of alternatives methodology of Guevara and Ben-Akiva (2013) presenting a probability specification as

$$P_{in} = \frac{e^{V_{in}+lnG_j(e^V)+ln\pi(D_n)}}{\sum_j e^{V_jn+lnG_j(e^V)+ln\pi(D_n)}} \tag{5}$$

$$\pi(D_n) = \frac{J^*_{r(i)n}}{J_{r(i)n}} \tag{6}$$

where $J^*_{r(i)n}$ is the number of alternatives sampled from a stratum r and $J_{r(i)n}$ is the total number of alternatives in that stratum.

3.3 Implementation of sampling algorithm and sampling protocols

3.3.1 The sampling algorithm

Sampling of alternatives can be implemented via random sampling i.e. alternatives are sampled randomly out of the total choice set. However, this approach may not result in rerpesentative choice sets. The second alternative is via importance sampling. The researcher subsets the alternatives into strata based on a set of deterministic rules which are defined based on the details of the specific problem under investigation. The sample size of alternatives within each stratum is then decided in order to give higher chance for some alternatives to be sampled. It must be noted that each alternative can be assigned to one stratum only. A pseudo-algorithm for sampling alternatives is as follows:

1. Generate a number of M (m_1, m_2,...,M) strata each with k_m sample size.
2. Validate the strata i.e. make sure (a) there are not empty strata and (b) the target sample size is not larger than the number of alternatives within each stratum.
3. Calculate the SC term as the ln of Equation 6.
4. For each individual sample alternatives within each stratum with a uniform probability.
5. If the chosen alternative has not been selected, it is added to the appropriate stratum and another alternative is removed.
6. If the chosen alternative is not selected, then add it deterministically to the appropriate stratum and remove another alternative from the same stratum.
7. Repeat the process for each individual

3.3.2 Sampling protocols

Two sampling protocols are examined in the current paper:

- Random sampling: This approach does not involve any process of stratification. The algorithm is generating a subset of all alternatives based on a target number of alternatives. Each alternative has an equal probability to be selected.
- Importance sampling: Importance sampling requires the generation of a number of strata based on a set of criteria. The rules for defining the strata and their sample size are defined by the researcher. Hence, some alternatives have a higher probability to be selected than others.

The procedure of importance sampling implemented in the paper followed two main principles:

1. Include a stratification approach that would follow to some extent the nesting structure of the CNL model presented by Beine *et al.* (2021).
2. Use the proportions of the selected countries to define the number of alternatives per stratum.

The strata generated based on the aforementioned principles are presented in Table 1 together with the implemented sample sizes for 20, 40 and, 60 sampled alternatives.

Table 1: Strata and sample sizes

Stratum	N	% chosen	Sample size (projected)			Sample size (implemented)		
			20	40	60	20	40	60
OECD & English speaking	5	65%	13	26	39	5	5	5
OECD & European	11	3%	1	1	2	1	1	2
OECD rest countries	3	3%	1	1	2	1	1	2
European rest countries	4	2%	0	1	1	-	1	1
Contiguous to India	8	3%	1	1	2	1	1	2
Other countries	53	24%	4	10	14	12	31	48

In Table 1, the percentages of countries chosen in each of the strata is presented (only including respondents who stated an aspiration to move). Hence, even though the stratum of OECD & English speaking countries consisted of five countries only, it was selected in

65% of the times mainly due to the presence of USA and UK. Based on these numbers, the sample size of this stratum should be the highest in the sampling process. However, given that only five countries were available, the sample size was constrained to that number. The extra alternatives were assigned to the other countries group given that all other strata had very low frequencies.

3.3.3 Testing parameter equivalence

The equivalence of parameters between the original model that considered the full choice set and each of the models using sampled alternatives was investigated using the t-test of individual parameter equivalence (Galbraith and Hensher, 1982). For each of the model parameters it was calculated the term

$$t_{diff} = \frac{\beta_k - \beta_{k*}}{\sqrt{\sigma_k^2 + \sigma_{k*}^2}} \tag{7}$$

where β_k and β_{k*} represent the parameter estimates of the full sample model and models with sampled alternatives while σ_k and σ_{k*} are their standard errors. If $|t_{diff}| > 1.96$ then the parameters from models with sampled alternatives are not transferable to the full sample model.

4 Preliminary results

4.1 MNL model

The parameter estimates of the MNL model are presented in Section A.1 of the Appendix. In almost all cases, the parameter signs were retained regardless of the sample size of alternatives. The transferability results presented in Table 2 do not show any clear patterns regarding which sampling protocol produced less biased parameter estimates. However, it

must be mentioned that in all cases, no significant results were observed ($|t_{diff}| > 1.96$) hence it is likely that for the MNL model, even smaller samples of alternatives can produce reasonably good estimates. However, given that the parameters were tested with only one set of sampled alternatives per sample size no definite conclusions can be reached; additional testing is required with more samples to obtain more robust results.

Table 2: Transferability statistics for the MNL model

	20 alternatives		40 alternatives		60 alternatives	
	Random	Importance	Random	Importance	Random	Importance
Age over 65	0.138	0.727	-0.568	-0.381	0.464	0.163
Age under 65	-0.299	-1.033	1.184	-0.223	0.071	0.353
Male × HS	0.048	0.737	-0.229	-0.382	0.260	0.015
Male × LS	0.000	1.398	-0.287	-0.157	0.385	0.177
Male × MS	-0.561	0.274	1.410	-0.428	0.296	0.601
High skilled (HS)	0.592	0.366	-0.628	-0.130	0.131	-0.132
Medium skilled (MS)	0.697	-0.281	-0.596	-0.154	-0.153	-0.309
Low skilled (LS)	0.281	0.148	-0.764	-0.157	0.000	-0.316
Large city	1.051	-0.055	-0.154	0.259	0.396	-0.228
Log of income orig.	-1.063	-0.423	-0.199	0.495	-0.216	-0.129
More than 2 children	0.700	-1.178	0.746	0.126	-0.032	0.072
Network × HS	-0.579	-0.674	-0.430	0.415	-0.498	-0.319
Network × LS	0.717	-0.255	0.134	-0.310	-0.354	0.071
Network × MS	0.836	0.049	0.349	0.467	-0.121	-0.186
No child	0.744	-0.615	-0.026	-0.539	-0.402	0.097
Log of inc. at dest × HS	0.295	0.226	-0.148	0.038	0.000	0.000
Log of inc. at dest × LS	0.126	0.000	-0.130	0.000	0.000	0.000
Log of inc. at dest × MS	0.113	0.000	-0.176	0.000	0.059	0.000
Log of diaspora × HS	0.060	-0.180	0.000	-0.077	0.015	-0.015
Log of diaspora × LS	0.235	0.174	0.039	-0.039	-0.020	0.000
Log of diaspora × MS	-0.054	-0.219	0.112	-0.130	-0.094	0.019
Log of distance × HS	0.043	-0.174	-0.199	-0.260	0.074	-0.092
Log of distance × LS	0.206	-0.382	-0.085	-0.174	0.086	-0.043
Log of distance × MS	0.159	-0.034	-0.109	-0.142	0.088	-0.067
Log of population	0.495	0.155	0.104	0.088	0.087	0.035
Religious proximity × HS	-0.290	-0.468	-0.098	-0.735	-0.227	0.119
Religious proximity × LS	-0.239	-0.408	-0.190	-0.305	-0.474	0.000
Religious proximity × MS	-0.102	0.390	-0.149	-0.248	0.122	-0.042
$\delta_{Contiguous}$	0.193	-0.260	-0.240	-0.176	0.088	-0.066
$\delta_{English}$	0.044	-0.044	-0.090	0.045	0.046	0.000
$\delta_{European}$	0.193	-0.042	-0.252	-0.084	0.070	-0.021
δ_{OECD}	-0.024	0.044	0.099	0.089	-0.024	-0.017
δ_{Other}	0.314	-0.081	-0.027	-0.022	0.038	-0.036

4.2 NL model

The parameter estimates of the NL model are presented in Section A.2 of the Appendix. The nested logit specification included a two-nest structure. In particular, a nest included all *stayers* hence India and another nest all the *Foreign* destinations. A very interesting finding is that for 20 alternatives it was not possible to estimate the $\mu_{Foreign}$ nest parameter. Although this finding is related to the specific draw of samples (and possibly will not happen in another draw of sampled alternatives) it still denotes that in lower sample sizes, random sampling may not be suffucient and more sophisticated sampling protocols are required. The transferability results are presented in Table 3. Unlike the MNL model, transferability is not achieved for several parameters for 20 and 40 sampled alternatives (marked in bold). Moreover, random sampling fails more often to produce transferable parameters compared to importance sampling. Some parameters such as $\delta_{English}$ and *log of population* are consistently not being estimated well.

5 Conclusion and next steps

This paper presents the preliminary results of a sampling of alternatives example on migration aspiration discrete choice models. These initial results suggest that model complexity increases the need for more alternatives in order reduce the bias compared to the full sample results. Especially, random sampling performed less efficiently than the importance sampling for NL model which suggests the need for the development of more elaborated sampling protocols. However, it must be highlighted that the results presented in this work are based on single draws of samples; more samples of the same sizes must be generated and evaluated in order to obtain better and more accurate insights with respect to the impact of sample size on the bias of parameter estimates. Building on the current results, the next steps involve:

- Incluion of additional countries in the model specification as the original work of Beine *et al.* (2021) considered 85 alternatives.
- Re-estimation of the MNL and NL models on more sampled data sets
- Estimation of the CNL model and implementation of sampling of alternatives following Guevara and Ben-Akiva (2013) and Lai and Bierlaire (2015)
- Definition of and testing of additional sampling protocols for sampling alternatives

Table 3: Transferability statistics for the NL model

	20		40		60	
	Random	Importance	Random	Importance	Random	Importance
Age over 65	0.546	-0.355	-0.249	0.229	-0.181	0.000
Age under 65	-0.364	-1.301	0.674	0.199	-0.093	0.000
Male × HS	-0.526	-0.769	0.314	0.658	0.183	-0.280
Male × LS	-0.465	-0.599	-0.886	0.000	-0.690	0.060
Male × MS	1.624	-0.190	0.358	0.032	-0.209	-0.089
High skilled (HS)	**-2.948**	-1.232	**-3.206**	-1.934	-1.891	-0.930
Medium skilled (MS)	**-3.032**	-1.174	**-3.137**	-1.886	-1.774	-0.869
Low skilled (LS)	**-3.237**	-1.446	**-3.385**	**-2.013**	-1.872	-0.972
Large city	-0.964	0.402	-0.670	-0.646	-0.455	-0.089
Log of income orig.	**-2.553**	-0.835	0.444	0.512	0.348	0.102
More than 2 children	-1.707	0.386	0.430	0.561	0.675	0.241
Network × HS	-0.245	0.123	0.477	-0.104	0.064	-0.005
Network × LS	-0.027	0.924	0.410	-0.557	-0.664	0.102
Network × MS	0.033	-0.262	-1.574	0.203	0.341	-0.019
No child	-0.796	0.483	-0.743	-0.568	-0.314	-0.076
Log of inc. at dest × HS	**-2.727**	-1.104	**-2.465**	-1.770	-1.562	-0.968
Log of inc. at dest × LS	**-3.016**	-0.972	**-2.883**	-1.816	-1.647	-0.859
Log of inc. at dest × MS	**-2.817**	-0.963	**-2.867**	**-1.961**	-1.687	-1.019
Log of diaspora × HS	-1.027	-1.040	-1.122	-1.128	-0.808	-0.830
Log of diaspora × LS	**-2.243**	-1.507	**-2.371**	**-2.104**	-1.581	-1.334
Log of diaspora × MS	**-2.680**	-1.483	**-2.600**	**-2.003**	-1.618	-1.198
Log of distance × HS	1.462	-0.295	1.600	0.250	1.089	-0.225
Log of distance × LS	**2.807**	0.521	**2.669**	1.181	1.605	0.505
Log of distance × MS	**2.175**	0.032	**2.053**	0.740	1.319	0.169
Log of population	**-3.173**	-1.036	**-3.230**	**-1.959**	-1.869	-0.959
Religious proximity × HS	-0.873	-0.494	-1.007	-0.904	-1.040	-0.832
Religious proximity × LS	**-2.180**	-0.953	-1.985	-1.647	-1.365	-0.934
Religious proximity × MS	**-2.372**	-0.548	**-2.482**	-1.691	-1.400	-0.877
$\delta_{Contiguous}$	1.507	0.441	1.513	-0.326	1.124	-0.209
$\delta_{English}$	**-3.032**	**2.380**	**-3.009**	**2.203**	-1.719	1.127
$\delta_{EUropean}$	0.812	-0.985	0.934	-1.143	0.674	-1.643
δ_{OECD}	0.484	**2.390**	0.379	0.972	0.287	1.169
δ_{Other}	-0.207	1.793	-0.233	**3.299**	-0.201	**2.730**
$\mu_{Foreign}$	**1.964**	1.180	1.821	1.578	1.338	1.009

The ultimate aim of this exercise is understanding which sampling protocols work best for the case of migration aspiration models. These will allow for the estimation of discrete choice models that allow for complex correlation structures, such as CNL, by reducing computational time.

6 References

Beine, M., M. Bierlaire and F. Docquier (2021) New York, Abu Dhabi, London or Stay at Home? Using a Cross-Nested Logit Model to Identify Complex Substitution Patterns in Migration.

Bekaert, E., I. Ruyssen and S. Salomone (2021) Domestic and international migration intentions in response to environmental stress: A global cross-country analysis, *Journal of Demographic Economics*, **87** (3) 383–436, ISSN 2054-0892. Publisher: Cambridge University Press.

Bertoli, S. and J. F.-H. Moraga (2013) Multilateral resistance to migration, *Journal of development economics*, **102**, 79–100, ISSN 0304-3878. Publisher: Elsevier.

Bertoli, S. and I. Ruyssen (2018) Networks and migrants' intended destination, *Journal of Economic Geography*, **18** (4) 705–728, ISSN 1468-2702. Publisher: Oxford University Press.

Bierlaire, M., D. Bolduc and D. McFadden (2008) The estimation of generalized extreme value models from choice-based samples, *Transportation Research Part B: Methodological*, **42** (4) 381–394, ISSN 0191-2615. Publisher: Elsevier.

Buggle, J. C., T. Mayer, S. Sakalli and M. Thoenig (2020) The Refugee's dilemma: evidence from Jewish migration out of Nazi Germany. Publisher: CEPR Discussion Paper No. DP15533.

Docquier, F., A. Tansel and R. Turati (2020) Do emigrants self-select along cultural traits? Evidence from the MENA countries, *International Migration Review*, **54** (2) 388–422, ISSN 0197-9183. Publisher: Sage Publications Sage CA: Los Angeles, CA.

Galbraith, R. A. and D. A. Hensher (1982) Intra-metropolitan transferability of mode choice models, *Journal of Transport Economics and Policy*, 7–29, ISSN 0022-5258. Publisher: JSTOR.

Gubert, F. and J.-N. Senne (2016) Is the European Union attractive for potential migrants?: An investigation of migration intentions across the world. Publisher: OECD.

Guevara, C. A. and M. E. Ben-Akiva (2013) Sampling of alternatives in Multivariate

Extreme Value (MEV) models, *Transportation Research Part B: Methodological*, **48**, 31–52, ISSN 01912615.

Lai, X. and M. Bierlaire (2015) Specification of the cross-nested logit model with sampling of alternatives for route choice models, *Transportation Research Part B: Methodological*, **80**, 220–234, ISSN 01912615.

Langella, M. and A. Manning (2021) Income and the desire to migrate, ISSN 2042-2695. Publisher: Centre for Economic Performance, LSE.

Lovo, S. (2014) Potential migration and subjective well-being in Europe, *IZA Journal of Migration*, **3**, 1–18. Publisher: Springer.

McFadden, D. (1977) Modelling the choice of residential location.

Monras, J. (2018) Economic shocks and internal migration. Publisher: CEPR Discussion Paper No. DP12977.

Nerella, S. and C. R. Bhat (2004) Numerical analysis of effect of sampling of alternatives in discrete choice models, *Transportation Research Record*, **1894** (1) 11–19, ISSN 0361-1981. Publisher: SAGE Publications Sage CA: Los Angeles, CA.

Ortega, F. and G. Peri (2013) The effect of income and immigration policies on international migration, *Migration Studies*, **1** (1) 47–74, ISSN 2049-5838. Publisher: Oxford University Press.

Swait, J. and M. Ben-Akiva (1987) Incorporating random constraints in discrete models of choice set generation, *Transportation Research Part B: Methodological*, **21** (2) 91–102, ISSN 0191-2615. Publisher: Elsevier.

Tsoleridis, P., C. F. Choudhury and S. Hess (2022) Utilising activity space concepts to sampling of alternatives for mode and destination choice modelling of discretionary activities, *Journal of Choice Modelling*, **42**, 100336, ISSN 17555345.

A Parameter estimates

A.1 MNL model parameter estimates

	Full choice set		20 alternatives				40 alternatives				60 alternatives			
			Random		Importance		Random		Importance		Random		Importance	
	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test
Age over 65	-0.0435	-1.87	-0.0512	-1.01	-0.0736	-2.15	-0.0155	-0.357	-0.0269	-0.729	-0.059	-2.46	-0.0488	-2.15
Age under 65	0.0381	13.9	0.04	6.98	0.0454	6.97	0.0325	8.43	0.0391	11	0.0378	11.8	0.0367	12.8
Male × HS	-0.322	-2.34	-0.337	-1.2	-0.558	-1.93	-0.265	-1.28	-0.238	-1.39	-0.377	-2.34	-0.325	-2.26
Male × LS	-1.1	-9.5	-1.1	-4.92	-1.55	-5.16	-1.04	-5.97	-1.07	-7.05	-1.17	-8.35	-1.13	-9.15
Male × MS	-0.605	-6.48	-0.486	-2.55	-0.666	-3.3	-0.847	-5.88	-0.54	-4.51	-0.648	-5.82	-0.688	-6.76
High skilled (HS)	12.3	22.9	11.8	18.1	12	19.4	12.8	21.8	12.4	22.5	12.2	22.5	12.4	23.1
Medium skilled (MS)	13.3	29.4	12.8	23	13.5	24.5	13.7	27.6	13.4	28.6	13.4	28.5	13.5	29.2
Low skilled (LS)	14.1	31.8	13.9	25	14	27.5	14.6	30.3	14.2	31.2	14.1	30.9	14.3	31.7
Large city	-0.355	-5.13	-0.515	-3.8	-0.346	-2.33	-0.336	-3.3	-0.384	-4.36	-0.397	-4.94	-0.332	-4.51
Log of income orig.	-0.0742	-1.83	0.0258	0.304	-0.0303	-0.317	-0.0599	-1.01	-0.107	-2.04	-0.0607	-1.28	-0.0666	-1.56
More than 2 children	0.0767	0.844	-0.0641	-0.357	0.346	1.65	-0.0419	-0.321	0.058	0.493	0.0812	0.764	0.0672	0.701
Network × HS	-0.889	-5.88	-0.689	-2.22	-0.66	-2.17	-0.772	-3.41	-0.987	-5.45	-0.772	-4.3	-0.819	-5.15
Network × LS	-0.896	-6.55	-1.1	-4.41	-0.807	-2.51	-0.929	-4.56	-0.825	-4.49	-0.82	-4.96	-0.91	-6.33
Network × MS	-0.748	-7.02	-0.946	-4.47	-0.76	-3.41	-0.812	-5.43	-0.828	-6.17	-0.728	-5.76	-0.719	-6.29
No child	-0.138	-2.18	-0.247	-1.87	-0.0453	-0.331	-0.135	-1.41	-0.0819	-0.992	-0.0985	-1.31	-0.147	-2.18
Log of inc. at dest × HS	1.37	7.35	1.29	6.54	1.31	6.93	1.41	7.21	1.36	7.48	1.37	7.39	1.37	7.33
Log of inc. at dest × LS	1.09	10.2	1.07	9.13	1.09	9.46	1.11	9.96	1.09	9.95	1.09	10.1	1.09	10.1
Log of inc. at dest × MS	1.17	9.78	1.15	8.78	1.17	9.19	1.2	9.83	1.17	9.67	1.16	9.68	1.17	9.79
Log of diaspora × HS	0.107	2.3	0.103	2.19	0.119	2.49	0.107	2.26	0.112	2.45	0.106	2.27	0.108	2.32
Log of diaspora × LS	0.241	6.71	0.229	6.31	0.232	6.22	0.239	6.56	0.243	6.59	0.242	6.82	0.241	6.75
Log of diaspora × MS	0.299	7.95	0.302	7.41	0.311	7.78	0.293	7.63	0.306	7.95	0.304	8	0.298	7.92
Log of distance × HS	-0.645	-3.82	-0.656	-3.45	-0.603	-3.5	-0.596	-3.34	-0.583	-3.47	-0.663	-3.83	-0.623	-3.67
Log of distance × LS	-1.28	-7.88	-1.33	-7.35	-1.19	-6.96	-1.26	-7.4	-1.24	-7.61	-1.3	-7.75	-1.27	-7.66
Log of distance × MS	-0.937	-5.96	-0.975	-5.43	-0.929	-5.39	-0.912	-5.48	-0.905	-5.62	-0.957	-5.88	-0.922	-5.8

	Full choice set		20 alternatives				40 alternatives				60 alternatives			
			Random		Importance		Random		Importance		Random		Importance	
	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test
Log of population	0.73	18.2	0.7	15.4	0.721	17.2	0.724	17.3	0.725	17.9	0.725	17.7	0.728	18.1
Religious proximity × HS	0.979	2.97	1.16	2.19	1.24	2.76	1.03	2.54	1.32	4.04	1.09	3.03	0.923	2.76
Religious proximity × LS	1.39	7.89	1.46	6.26	1.51	6.41	1.44	7.36	1.47	7.57	1.51	8.3	1.39	7.67
Religious proximity × MS	1.44	8.69	1.47	6.07	1.33	5.82	1.48	7.02	1.5	8.48	1.41	7.73	1.45	8.48
$\delta_{Contiguous}$	-1.2	-3.74	-1.29	-3.81	-1.08	-3.25	-1.09	-3.33	-1.12	-3.48	-1.24	-3.81	-1.17	-3.64
$\delta_{English}$	2	12.9	1.99	11.9	2.01	12.3	2.02	12.7	1.99	12.6	1.99	12.8	2	12.9
$\delta_{EUropean}$	-0.183	-1.81	-0.212	-1.91	-0.177	-1.75	-0.146	-1.37	-0.171	-1.69	-0.193	-1.9	-0.18	-1.77
δ_{OECD}	-0.166	-0.802	-0.159	-0.746	-0.179	-0.842	-0.195	-0.936	-0.192	-0.935	-0.159	-0.768	-0.161	-0.781
δ_{Other}	0.139	0.542	0.0224	0.0835	0.169	0.636	0.149	0.566	0.147	0.576	0.125	0.484	0.152	0.594
Final LL	-8077.983		-3137.369		-3789.279		-4481.91		-5917.426		-6509.823		-7430.039	
AIC	16221.97		6340.738		7644.559		9829.821		11900.85		13085.65		14926.08	
BIC	16498.79		6617.567		7921.388		10106.65		12177.68		13362.48		15202.91	

A.2 NL model parameter estimates

	Full choice set		20				40				60			
			Random		Importance		Random		Importance		Random		Importance	
	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test	Value	Rob. t-test
Age over 65	-0.0426	-1.84	-0.0632	-2.12	-0.0237	-0.495	-0.0321	-0.91	-0.0513	-1.7	-0.0362	-1.35	-0.0426	-1.67
Age under 65	0.0383	14	0.0407	6.79	0.0472	7.53	0.035	8.62	0.0374	10.4	0.0387	11.6	0.0383	13.1
Male × HS	-0.318	-2.32	-0.16	-0.599	-0.0884	-0.333	-0.393	-2.01	-0.466	-2.61	-0.358	-2.11	-0.262	-1.8
Male × LS	-1.1	-9.51	-0.969	-3.77	-0.942	-3.97	-0.926	-5.83	-1.1	-7.23	-0.977	-7.2	-1.11	-9.19
Male × MS	-0.597	-6.43	-0.98	-4.52	-0.552	-2.53	-0.657	-4.7	-0.602	-4.9	-0.567	-5.2	-0.585	-5.94
High skilled (HS)	6.03	4.7	12	7.66	8.39	5.9	11.3	11	9.58	7.3	9.19	8.58	7.83	5.4
Medium skilled (MS)	7.33	6.07	12.9	9.32	9.45	7.04	12.1	13.1	10.5	8.98	10.1	10.2	8.9	6.62
Low skilled (LS)	7.09	4.99	14	8.79	10.1	6.64	13.1	12.3	11.1	7.95	10.5	9.21	9.15	5.82
Large city	-0.36	-5.21	-0.202	-1.36	-0.426	-2.86	-0.278	-2.75	-0.286	-3.13	-0.311	-3.76	-0.351	-4.78
Log of income orig.	-0.0712	-1.76	0.157	1.97	0.0103	0.116	-0.103	-1.74	-0.105	-2.01	-0.0932	-1.92	-0.0772	-1.81
More than 2 children	0.0659	0.725	0.459	2.17	-0.0178	-0.0905	-0.00081	-0.00644	-0.0165	-0.143	-0.0278	-0.265	0.0342	0.359
Network × HS	-0.91	-6.06	-0.828	-2.77	-0.95	-3.31	-1.03	-5.1	-0.885	-4.71	-0.925	-5.14	-0.909	-5.7
Network × LS	-0.915	-6.71	-0.906	-2.93	-1.19	-4.5	-1.01	-5.4	-0.787	-4.25	-0.77	-4.52	-0.935	-6.6
Network × MS	-0.778	-7.32	-0.786	-3.55	-0.706	-2.78	-0.463	-2.73	-0.813	-6	-0.833	-6.87	-0.775	-6.9
No child	-0.139	-2.2	-0.0217	-0.163	-0.214	-1.51	-0.0557	-0.601	-0.0794	-0.948	-0.108	-1.42	-0.132	-1.96
Log of inc. at dest × HS	0.553	3.06	1.46	5.23	0.867	3.95	1.22	6.06	1.06	4.77	0.955	5.21	0.833	3.69
Log of inc. at dest × LS	0.438	3.18	1.09	6.54	0.64	4.11	0.979	7.68	0.801	5.53	0.744	5.97	0.623	3.76
Log of inc. at dest × MS	0.47	3.23	1.14	6.06	0.688	3.97	1.05	7.47	0.895	5.57	0.807	5.9	0.702	4.01
Log of diaspora × HS	0.0408	1.85	0.0983	1.91	0.0891	2.18	0.0965	2.17	0.0952	2.22	0.073	2.2	0.0749	2.16
Log of diaspora × LS	0.0948	3.11	0.227	4.5	0.176	3.96	0.21	5.55	0.207	4.73	0.166	5.01	0.162	4.04
Log of diaspora × MS	0.118	3.05	0.301	5.35	0.213	4.17	0.268	6.26	0.248	4.76	0.206	5.39	0.194	3.86
Log of distance × HS	-0.245	-2.42	-0.575	-2.85	-0.197	-1.55	-0.554	-3.37	-0.289	-2.01	-0.427	-3.21	-0.21	-1.78
Log of distance × LS	-0.5	-3.14	-1.29	-5.56	-0.623	-3.57	-1.14	-6.36	-0.787	-4.29	-0.869	-5.24	-0.618	-3.62
Log of distance × MS	-0.361	-2.94	-0.885	-4.27	-0.367	-2.66	-0.793	-4.64	-0.508	-3.25	-0.611	-4.23	-0.392	-2.88
Log of population	0.289	3.33	0.695	7.39	0.421	4.51	0.644	9.55	0.53	6.08	0.502	6.8	0.415	4.21
Religious proximity × HS	0.594	2.74	1	2.43	0.811	2.12	1.03	2.75	0.94	2.98	0.957	3.5	0.872	3.43
Religious proximity × LS	0.712	3.35	1.46	5.42	1.01	4.4	1.3	6.3	1.2	5.81	1.09	6.14	0.986	4.88
Religious proximity × MS	0.71	3.12	1.63	5.19	0.883	4.04	1.5	6.74	1.25	5.58	1.11	6.43	0.986	4.54
$\delta_{Contiguous}$	-0.484	-2.56	-1.11	-3	-0.613	-2.75	-1.04	-3.3	-0.383	-1.56	-0.837	-3.34	-0.427	-2.17
$\delta_{English}$	0.792	3.2	1.98	6.52	0.131	1.04	1.78	8.25	0.168	1.22	1.36	6.21	0.47	3.29
$\delta_{European}$	-0.065	-1.45	-0.166	-1.43	0.0109	0.174	-0.162	-1.73	0.034	0.459	-0.122	-1.7	0.0557	0.957
δ_{OECD}	-0.0796	-0.94	-0.193	-0.883	-0.544	-3.11	-0.158	-0.837	-0.255	-1.6	-0.127	-0.894	-0.27	-1.94
δ_{Other}	0.0383	0.371	0.0984	0.363	-0.313	-1.88	0.0979	0.419	-0.778	-3.46	0.0794	0.449	-0.566	-2.89
$\mu_{Foreign}$	2.54	3.29	1	7.24	1.55	4.72	1.12	10.1	1.28	6.26	1.47	7.08	1.67	4.35
Final LL	-8068.546		-2991.298		-3738.364		-5007.984		-5915.427		-6421.921		-7443.139	
AIC	16205.09		6050.596		7544.728		10083.97		11898.85		12911.84		14954.28	
BIC	16490.31		6335.813		7829.946		10369.19		12184.07		13197.06		15239.5	