Looking for opportunities – On aggregation in random utility models for migration*

Liesbeth Colen
DARE, University of Göttingen

Damiaan Persyn
Thünen Institute of Rural Economics
DARE, University of Göttingen

First version: November 11, 2021

This version: July 24, 2022

Latest version: https://drive.google.com/file/d/1o1skFAzEl409pCEwMw7GQP7QEtpQsbmv

Abstract

This paper considers a random utility model (RUM) in which migrants view locations as aggregates of large numbers of alternatives from which they can freely choose. The best alternative is more likely to be found in a location where they are many and diverse. This predicted effect of size and dispersion contrasts with models considering regions or countries as atomistic units of choice, or where the outcome obtained within a location is stochastic and more dispersion implies more uncertainty. The coefficient on size equals 1 in an ideally specified RUM model including all relevant control variables and appropriate nesting of locations such that residual correlation between alternatives within locations is small. Only then intuitive spatial properties hold: there is zero predicted net migration between otherwise similar regions of different size, and migration flows scale proportionally when aggregating locations. Imposing proportional scaling also constrains how measures of size corresponding to distinct sets of alternatives (e.g. the number of houses and jobs in a location) should be combined. Lastly, assuming normally distributed returns from individual alternatives within locations, the coefficient on the variance should be close to 0.5 in the suggested framework. The approach is showcased and key predictions are tested in a study of internal migration and urbanisation in Ethiopia.

*email: *damiaan.persyn@uni-goettingen.de*. We would like to thank Tom Bundervoet and Astewale Melaku for their kind help and comments, and Simone Bertoli and the seminar participants at the chair of Agricultural policy at the DARE, and of the GfR meeting in Schwerin for helpful comments. Any remaining errors are ours.

1 Introduction

This papers considers prospective migrants who view locations as aggregates of fundamental units of choice. Depending on the context, such a unit of choice could be a job, a dwelling, a partner, a piece of arable land, etc. We call these 'alternatives', or 'opportunities'.

Some well known models for migration consider locations as aggregates, but typically these are not aggregates of alternatives between which migrants can freely choose. In Harris and Todaro (1970) or Katz and Stark (1986) locations are viewed as lotteries over alternatives. Migrants choose locations depending on their ex-ante *expected utility* and draw an alternative on arrival. In Borjas (1987), prospective migrants are first assigned an outcome from a conditional distribution in every location, and then choose the best location given these draws.

In this paper, we apply a less known extension of the nested logit model of McFadden (1977) to migration.[1] In this framework migrants have idiosyncratic preferences over all alternatives in all locations, and with full information choose their best alternative. The probability that a location is chosen (contains the single best alternative) depends on the *expected maximum utility* the location provides, which increases with the number alternatives it contains (the destination 'size'), and with higher dispersion in the utility derived from the opportunities within a location. It is clearly context dependent whether a framework assuming stochastic outcomes in destinations as in Harris and Todaro (1970) and Katz and Stark (1986) is suited, or rather one assuming perfect information and free choice as considered in this paper.

In the next paragraphs we summarise four insights obtained from the proposed framework, as well as some testable predictions, and point out the contributions of this paper.

A first result from considering locations as aggregates of choice alternatives following McFadden (1977) is that the utility from choosing a location increases log-linearly in the number of alternatives it contains. When deriving a gravity equation from the discrete choice framework, the number of opportunities available in a destination country or region shows up as an attractive factor, serving as the size or mass variable for the destination. A first contribution of this paper is to point to the logic and importance of including a size variable for the destination in analyses of migration. Current economic research on migration (see Beine, Bertoli, and Fernández-Huertas Moraga, 2016, for an overview) does not explicitly consider choice within locations. In these models size or dispersion are missing or added

[1] See section 8 'Aggregation of alternatives and the treatment of similarities' in McFadden (1977), available here: https://drive.google.com/file/d/1QJAH2jR0HWGYy41qgTHItUgJ2UESFsue/view?usp=sharing, and also Lerman (1975) pages 155-164 for aggregation in a multinomial context. Kanaroglou and Ferguson (1996) relax some of the underlying assumptions.

in an ad-hoc fashion. We point out some well known studies in this literature that may suffer from bias due to not controlling adequately for the size of the opportunity set in the destinations.

Second, in the suggested framework the coefficient associated with the number of underlying alternatives in a location (i.e. the coefficient on destination size) reflects the dissimilarity of the alternatives within the location. In an ideal model including the most relevant variables affecting migration and implementing appropriate nesting of destinations[2], residual correlation in the utility derived from different opportunities within countries should be small, opportunities should be perceived as dissimilar, and the mass variable should therefore have an associated coefficient close to 1.[3] A coefficient on size significantly different from 1 can point to a mass variable that does not proxy well for the number or the type of opportunities in a destination, or point to significant residual correlation between opportunities within destinations, which would suggest that an important control variable is missing or additional nesting of destinations should be considered. This possible interpretation of the coefficient on size is hardly recognised in current economic research on migration, or contributions using gravity equations in various contexts within economics. A second contribution of this paper is to remind modellers considering spatial aggregates such as regions or countries of this interpretation of the coefficient on size offered by the micro-foundation using random utility theory.

Third, a coefficient of 1 on destination size relates to several spatial properties of the model: (1) As is well known in trip-distribution analysis since at least Daly (1982), a coefficient of 1 on destination size is required to make predicted flows independent from the level of spatial aggregation of destinations, such that the predicted migration flow to a country equals the predicted migration flows to its constituent regions and vice-versa. (2) Again following Daly (1982), if migrants are looking for combinations of opportunities (for example a job and a dwelling), and if predicted migration flows are to scale proportionally when considering aggregates of destinations, then the corresponding size variables should be combined (for example in a weighted index), entering the gravity equation as a single mass variable for

[2]Some key contributions in this literature such as Ortega and Peri (2013), Beine, Bourgeon, and Bricongne (2019) or Beine, Bierlaire, and Docquier (2021) consider nests of locations. This paper suggests that while this nesting of locations is highly relevant, it is insightful to additionally consider locations themselves as aggregates (nests) of many opportunities as in Section 8 of McFadden (1977).

[3]See also Train (2002) who notes that 'It is important to realize that the independence assumption is not as restrictive as it might at first seem, and in fact can be interpreted as a natural outcome of a well-specified model. [...] In a deep sense, the ultimate goal of the researcher is to represent utility so well that the only remaining aspects constitute simply white noise. That is: the goal is to specify utility well enough that a logit model [for which the dissimilarity parameter equals 1] is appropriate. Seen in this way, the logit model is the ideal rather than a restriction.'

the destination with an associated coefficient equal to 1. We implement the aggregation of multiple mass variables in the empirical application. To the best of my knowledge, this consistent combination of multiple mass variables has never been considered in the context of migration, or related frameworks such as trade gravity equations. (3) We show that a coefficient of 1 on size is required to have a spatial equilibrium with zero net migration between locations of different size with similar opportunities. It is intuitive that there exist spatial equilibria between regions or countries with similar opportunities, even if the number of these opportunities is very different, such as between two similar EU member states of different size (say Denmark and Germany) or a rural area within a country and a much larger aggregate of similar areas in the same country. This intuition does not hold in analyses omitting a mass variable, or when the coefficient on mass differs significantly from 1. These three properties of the coefficient on mass or size seem unknown in the migration literature.

Fourth, the expected maximum utility from choosing a destination, and therefore the expected migration flow to this destination, increases with the dispersion between the opportunities within the destination. Following McFadden (1977), for normally distributed opportunities, if the coefficient on size is close to 1, theory predicts a coefficient on the variance between opportunities of 0.5. This attractive property of dispersion contrasts sharply with models of migration where a prospective migrant draws an alternative within destinations under uncertainty and evaluates the expected utility in a destination such that higher dispersion in outcomes implies lower expected utility in presence of risk aversion as in Katz and Stark (1986). The attractiveness of variation in opportunities follows directly from the theory of random utility maximisation when considering locations as aggregates of units of choice as described in McFadden (1977), but seems little known or applied. It may offer an unexplored alternative explanation for the observed attractiveness of cities to migrants, in spite of high inequality in economic outcomes in cities and high unemployment rates; an issue studied by economists since at least Harris and Todaro (1970).

We apply the described methodology in a study of interregional migration in Ethiopia. Ethiopia provides an interesting setting to test the suggested framework because information flows are likely hampered over larger distances and across regional borders, given the significant ethno-linguistic heterogeneity within the country. It is found that dispersion in outcomes within a location is attractive for the current location, less attractive for neighbouring locations, and even less attractive when crossing ethno-linguistic regional borders. As information becomes less readily available over larger distances, a traditional framework with stochastic outcomes and risk-aversion may be increasingly relevant, rather than the

model perfect information and free choice considered in this paper.[4]

To control for unobserved differences between migrants and non-migrants, the familiar type of nesting of locations is used distinguishing between the own region and all other possible destinations as in Ortega and Peri (2013) and Beine, Bourgeon, and Bricongne (2019). The proposed framework is implemented in an additional lower level of nesting, which is implicit, as in section 8 of McFadden (1977). Different types of opportunities (different mass variables for destinations) are considered combined in a single index as in Daly (1982): population, the number of houses with running water, and the number of jobs with paid earnings.

Existing contributions that are closest to this work are Kanaroglou and Ferguson (1996) which also discuss the framework of McFadden (1977), and Ferguson and Kanaroglou (1997) who apply it to internal migration in Canada. These insightful contributions have received little attention. Relative to Kanaroglou and Ferguson (1996), we add by considering the predictions and implications related to the coefficients on size and dispersion. Ferguson and Kanaroglou (1997) take the number of census areas per province as the proxy for the province size. However, given that these areas are of very different size in terms of GDP or population, their number within a province may have little relevance to migrants, and the number of individuals, houses and jobs within locations we use in this paper may be better proxies of the size of the opportunity sets considered by migrants. Another important difference is that we do not to rely on the effect of dispersion for the identification of the dissimilarity parameter, given that the effect of dispersion likely to depends on information availability. We obtain results that are much more in line with theory compared to Ferguson and Kanaroglou (1997), whose estimates often violate the theoretical framework.

The remainder of this paper is organised as follows: Section 2 presents a nested logit framework for migration where locations are viewed as aggregates of opportunities. Section 3 considers the link with the existing literature. Section 4 considers the application to internal migration in Ethiopia. Section 5 concludes.

2 Theory

2.1 A nested logit model with locations nesting choice alternatives

An individual i living in location o considers different choice alternatives f, for example jobs. Each alternative belongs to a single location $d \in D$. Write F_d for the set of alternatives

[4] See also Bertoli, Moraga, and Guichard (2020) on how costly information acquisition can shape migration decisions.

in d. Choosing an alternative or opportunity $f \in F_d$ implies living in destination d, which involves migrating if $o \neq d$.

Starting from the notation of Beine, Bertoli, and Fernández-Huertas Moraga (2016), utility U of individual i is assumed to depend on an index of observables at the destination w_d, on the bilateral cost c_{od} of moving from o to d, and on factors z_{fd} specific to the chosen alternative f within the destination. As in Cardell (1997); Berry (1994) and Ortega and Peri (2013), also the unobserved part of individual utility is split in a destination specific part μ_{di}, and an opportunity-specific part ϵ_{ofi} such that

$$U_{ofi} = w_d + z_{fd} - c_{od} + (1 - \lambda_d)\mu_{di} + \lambda_d \epsilon_{ofi}. \tag{1}$$

All elements in the utility function are known by the individual. The unobserved (to the econometrician) part of utility which is shared among all opportunities within a destination for an individual μ_{di} is iid extreme value. The fully idiosyncratic part which also varies between opportunities ϵ_{ofi} is distributed as the unique random variable ensuring that also the joint error term $(1-\lambda_d)\mu_{di} + \lambda_d \epsilon_{ofi}$ is extreme value distributed. The 'dissimilarity parameter' λ_d governs the correlation between the unobserved part of utility for individuals between opportunities within destinations. A low value of λ_d implies that individuals perceive the opportunities in a destination as similar, increasing the role of the observed opportunity specific characteristics z_{fd} in the choice between opportunities within a given destination.

Given the assumptions on the error terms, the probability P_f that an alternative f within the opportunity set F_d of destination d is chosen has a closed form solution as shown by McFadden (1977). Consider the decomposition $P_f = P_d \cdot P_{f|d}$ which provides the following convenient and well-known representation of the nested logit model:

$$P_f = P_d \cdot P_{f|d}$$

$$P_d = \frac{\exp(w_d - c_{od} + \lambda_d I_d)}{\sum_e \exp(w_e - c_{oe} + \lambda_e I_e)} \tag{2a}$$

$$I_d = \log \sum_{g \in F_d} \exp(z_{gd}/\lambda_d). \tag{2b}$$

$$P_{f|d} = \frac{\exp(z_{fd}/\lambda_d)}{\sum_{g \in F_d} \exp(z_{gd}/\lambda_d)}$$

2.2 Location choice

In migration analysis interest lies with the locations (countries, regions) and the probabilities that they are chosen, which is given by P_d given in (2a) and (2b). Information on the individual alternatives within locations may be unavailable, and there may simply be little interest in knowing exactly which alternative (which job, which dwelling, which partner) was chosen.

As can be seen from expressions (2a) and (2b), the presence of a set of choice alternatives within destinations implies that *the probability that a particular destination is chosen does not simply depend on the expected utility derived from the opportunities it contains*, as is assumed in many applications. It rather depends on the *expected maximum utility* that can be obtained from choosing an element from the set F_d:

$$V_d \equiv E[\max_{f \in F_d} U_{ofi}] = w_d - c_{od} + \lambda_d I_d = w_d - c_{od} + \lambda_d \log \sum_{f \in F_d} \exp(z_{fd}/\lambda_d).$$

Write z_d for the average of the z_{df} within location d . Then V_d can be written as

$$V_d = w_d - c_{od} + z_d + \lambda_d \log(N_d) + \lambda_d \log\left(\frac{1}{N_d} \sum_{f \in F_d} \exp\left(\frac{z_{fd} - z_d}{\lambda_d}\right)\right),$$

where N_d is the number of elements in F_d which can be simply called the size of d. The last term is always positive and increases with larger dispersion of the z_{fd} from their mean. *This shows that the utility from choosing a location d is increasing in the location size and the dispersion between the opportunities it contains.* Again following McFadden (1977), if the opportunities are many and iid normally distributed $z_{fd} \sim \mathcal{N}(z_d, \sigma_d^2)$ it follows that this expression almost surely converges to

$$V_d \overset{\text{a.s.}}{\cong} w_d - c_{od} + z_d + \lambda_d \log(N_d) + 0.5\frac{\sigma_d^2}{\lambda_d}, \tag{3}$$

and the probability that a location d is chosen therefore is

$$P_d = \frac{\exp(w_d - c_{od} + z_d + \lambda_d \log(N_d) + 0.5\frac{\sigma_d^2}{\lambda_d})}{\sum_e \exp(w_e - c_{oe} + z_e + \lambda_d \log(N_e) + 0.5\frac{\sigma_e^2}{\lambda_d})}.$$

The choice at the aggregate (location) level therefore is described by a standard logit model, where the utility V_d derived from choosing location d includes measures of size and dispersion between the opportunities contained in d. Apart from including appropriate control variables

in the discrete choice model (or a gravity model derived from it) at the aggregate (location) level, the lower level choice between alternatives can be left implicit. There is no need to explicitly model the presence or the choice between the alternatives within locations, unless there is an interest in understanding the choice between them.

Following the notation of Beine, Bertoli, and Fernández-Huertas Moraga (2016), define $y_d = exp(w_d)$; $q_d = exp(z_d)$; $\zeta_d = exp(\sigma_d^2)$ and $\phi_{od} = exp(-c_{od})$, moreover write m_{od} for the number of migrants from o to d and pop_o for the population in o. If the number of prospective migrants is large, the probability of migration to a destination P_d equals the share s_{od} of migrants from o to d relative to the local population $P_d = s_{od} = \frac{m_{od}}{pop_o}$. Migration flows then are described by the following gravity equation:

$$m_{od} = pop_o y_d q_d N_d^{\lambda_d} \zeta_d^{0.5/\lambda_d} \phi_{od} \frac{1}{\sum_e y_e q_e N_e^{\lambda_d} \zeta_d^{0.5/\lambda_d} \phi_{od}}. \tag{4}$$

Here y_d collects the influence of variables pertaining the country (climate, etc.), q_d pertains to characteristics of the opportunities (average wage, housing price level, etc.), N_d is the number or mass of opportunities (number of jobs or houses, arable land area, etc.[5]), and the associated parameters $0 \leq \lambda_d \leq 1$ reflect how independent the unobserved part of utility is between opportunities in each destination. ζ_d controls for dispersion between opportunities, and ϕ_{od} is an inverse measure of the cost of migration from o to d.

$P_d = s_{od} = m_{od}/pop_o$ are the odds of migration to d. Using the odds of staying in o as the reference, the log odds ratio is given by

$$\ln\left(\frac{s_{od}}{s_{oo}}\right) = \ln\left(\frac{m_{od}}{m_{oo}}\right) = \lambda_d \ln(N_d) - \lambda_d \ln(N_o) + 0.5\frac{\sigma_d^2}{\lambda_d} - 0.5\frac{\sigma_o^2}{\lambda_o} + w_d - w_o + z_d - z_o - (c_{od} - c_{oo}). \tag{5}$$

Equations (4) and (5) can be estimated using aggregate data using Poisson maximum likelihood and OLS respectively. Here pop_o and N_d are the mass variables for origin and destination, and dispersion between alternatives enters as an additional control variable. As argued in the next section, in an ideal setting the parameters λ_d should be close to 1. An older version of this paper considers the link between the gravity equation (4) and the gravity equation of Anderson and Wincoop (2003) as well as the single and doubly constrained models of Wilson (1967, 1970, 1971).

[5]The next sections considers how multiple mass variables can be considered jointly.

2.3 Three arguments to aim for $\lambda_d = 1$ in applications to migration

The ideal RUM model As equation (1) shows, $\lambda_d < 1$ is indicative of correlation in the residual part of utility between the individual opportunities within a location. As more relevant control variables are included in $w_d - c_{od}$, the residual part μ_d that is shared between alternatives within d becomes smaller, and residuals become increasingly idiosyncratic and uncorrelated between alternatives. Individuals then increasingly perceive alternatives as dissimilar. Therefore $\lambda_d = 1$ would hold in an ideally specified model including all relevant control variables at the location level. The model then collapses to a multinomial logit model[6] between individual alternatives.

There may exist correlation between opportunities that is hard to control for using observables, leading to $\lambda_d \ll 1$. If the factors causing the correlation are shared between subsets of locations, however, then higher levels of nesting can be considered, grouping locations as in for example Ortega and Peri (2013) or Beine, Bourgeon, and Bricongne (2019). Considering the probabilities and utilities conditional on having chosen a specific nest, correlation between the alternatives should again be small and λ therefore be closer to 1. Controlling for such unobserved common factors between subsets of locations may amount to including dummies in a regression such as in Beine, Bourgeon, and Bricongne (2019).

In applied work population is often used as the mass variable for both origin and destination. However, expressions (4) and (5) show it is important to distinguish between the number of potential migrants as the mass variable in the origin, and the mass variable for the destinations which should rather capture the multitude of alternatives for migrants within the destination. When the proxy chosen as the size variable for the destinations, for example the destination population or number of jobs, correlates only weakly with this number of alternatives, one would expect $\lambda \ll 1$.

In short, a coefficient on destination size significantly different from 1 may point to model misspecification . There may be important explanatory variables missing from the analysis leading to correlation between alternatives within locations; some subsets of locations may have common unobserved characteristics, which calls for an additional level of nesting of locations or the inclusion of dummy variables capturing these factors; or the proxy for destination size may not capture well what migrants are looking for.

Aggregation of migration flows Because migration is a spatial phenomenon, it is interesting to consider the relationship between λ_d and spatial properties implied by the model.

[6] See also (Train, 2002, p. 42-43). Anas (1983) discusses aggregation in such a multinominal logit framework. His resulting gravity equation is isomorphic to Anderson and Wincoop (2003) and the doubly constrained models of Wilson (1967, 1970, 1971).

One such property is that only for $\lambda_d = 1$ the migration flow to an aggregate such as a country equals the sum of the predicted flows to its constituents regions. This can be seen directly from the gravity equation (4), or taking ratio of the migration flows to two destinations k and l assuming that their size differs by a factor R such that $N_k = RN_l$. Collecting all other determinants of the migration flows in the factors b_k and b_l one obtains in logs

$$\log(m_k) - \log(m_l) = \log(b_k) - \log(b_l) + \lambda_k \log(R) + \lambda_k \log(N_l) - \lambda_l \log(N_l).$$

which shows that migration flows increase proportionally with destination size R only if $\lambda_k = 1$. Only in this case the predicted migration flow to a country consisting of R regions of size N equals the sum of the predicted flows to the individual regions.

Alternatively, consider equation (4) while assuming away migration costs[7] such that $\phi = exp(-c_{od}) = 1$ and assuming all destination specific factors apart from size are equal among locations and normalised such that $b_d = y_d q_d \zeta_d^{0.5/\lambda_d} = 1$. Consider an aggregate of locations $S \subset D$. Then the flow to an aggregate of size $\sum_{j \in S} N_j$ equals the sum of the flows to each of its constituents of size N_j considered separately if

$$pop_o \left(\sum_{j \in S} N_j \right)^{\lambda} \frac{1}{\left(\sum_{j \in S} N_j \right)^{\lambda} + \sum_{j \in D \setminus S} N_j^{\lambda}} = \sum_{j \in S} \left(pop_o N_j^{\lambda} \frac{1}{\sum_{j \in S} (N_j^{\lambda}) + \sum_{j \in D \setminus S} N_j^{\lambda}} \right)$$

which holds for $\lambda = 1$.

Since a failure of migration flows to scale proportionally when aggregating locations corresponds to $\lambda < 1$, it may indicate model misspecification as discussed above. In presence of unobserved factors common to subsets of locations, nesting of locations may be required to ensure $\lambda_d \approxeq 1$ and proportional scaling to hold within the nests.[8]

Trip analyses often impose perfect scaling of flows with size by assuming $\lambda = 1$. This also restricts how different size variables can be considered jointly. If migrants choose sets of alternatives, such as a combination of a job and a dwelling, Daly (1982) suggests combining

[7] In presence of migration costs $\lambda = 1$ is still required for migration to the aggregate to be the sum of the migration flows of its constituents, but the notation would have to keep track of the role of migration costs in the internal structure of the aggregate. This was considered in an earlier draft of this paper.

[8] Consider the textbook example of the choice between transport modes bus-car-bike. If each mode has a choice probability of 1/3, the choice probability of an aggregate of 2 alternatives scales proportionally, equalling 2/3. If we rather consider three buses of a different colour with choice probabilities of 1/9 each alongside car and bike (with probabilities 1/3), proportional scaling fails: if we consider aggregates of a bus of a specific colour and a car, the probability will be 1/9+1/3. By rather nesting the buses in a separate nest, probabilities again scale proportionally within the non-bus and bus nests.

the corresponding size variables in a single weighted index. Assuming homogeneous[9] opportunities, with different relevant size variables in the destination, $N_{1d}, N_{2d}, \ldots$ the utility from choosing destination d can be modelled as (compare to equation (3))

$$V_d = w_d - c_{od} + z_d + \lambda \log(N_{1d} + b_2 N_{2d} + \ldots).$$

The single index $N_d = N_{1d} + b_2 N_{2d} + \ldots$ would enter the gravity equation as the combined mass or size variable for the destination. The index weights b can be estimated from data. Combining multiple mass variables in a single index is compatible with proportional scaling if $\lambda = 1$.

Spatial equilibrium Consider the definition of spatial equilibrium used in quantitative spatial models, such as Behrens and Murata (2021) or Kline and Moretti (2014). The long run is considered with migration costs set to 0. Agent characteristics and preferences are assumed to be independent from the current location of residence, such that in spatial equilibrium the population share of a region equals its choice probability by the agents, or following the notation of Behrens and Murata (2021):

$$P_d = \frac{exp(V_d/\beta_d)}{\sum_e exp(V_e/\beta_e)} = \frac{pop_d}{\sum_e pop_e},$$

where β_d is the inverse of the scale parameter of the EV distributed residuals which was normalised in Section 2. Accounting for differences between locations in the number of underlying opportunities, for example through differences in the level of spatial aggregation between units, can be introduced by defining $V_d = V'_d + \beta'_d \log(N)$ where V'_d collects properties other than size, β'_d is the inverse of the scale parameter of residuals at the lower level between opportunities within locations such that $\beta_d > \beta'_d$ and $0 \leq \lambda_d = \beta'_d/\beta \leq 1$. The spatial equilibrium then is defined by

$$P_d = \frac{\exp(V'_d/\beta_d + \frac{\beta'_d}{\beta_d}\log(N_d))}{\sum_e \exp((V'_e/\beta_e + \frac{\beta'_e}{\beta_e}\log(N_e))} = \frac{N_d^{\lambda_d}\exp(V'_d/\beta_d))}{\sum_e N_d^{\lambda_d}\exp((V'_e/\beta_e)} = \frac{pop_d}{\sum_e pop_e}.$$

[9] In case of heterogeneous opportunities, the expression would additionally depend on the dispersion between the various opportunities. This is left for future research.

Now consider the ratio of the equilibrium population of two regions k and l of different size such that $N_k = RN_l$

$$\frac{(RN_l)^{\lambda_k} exp(V_k'/\beta_k)}{N_l^{\lambda_l} exp(V_l'/\beta_l)} = \frac{pop_m}{pop_n}.$$

With $\lambda_k = 1$, the long run population distribution scales proportionally when aggregating locations. If two spatial aggregates are different in size but are otherwise similar in terms of key variables which may affect migration such as for example population density, average wages, housing prices and the variance of the idiosyncratic residuals governed by β_i, intuitively a spatial aggregate twice as large should contain twice the population in equilibrium. In presence of agglomeration economies locations with a high population density are more attractive. This can best be modelled by including population density in V'. The framework then allows to decouple the concepts of size and density. Modelling size differences as suggested while imposing $\lambda = 1$ would ensure that an aggregate containing two similar cities with a similar population density will have an equilibrium population that is twice as large as an aggregate containing one such city, while simultaneously allowing cities to have different equilibrium populations compared to rural areas through the influence of population density included as a variable in V'.

As before with migration flows, if equilibrium population does not scale with location size ($\lambda \ll 1$), it may point to model misspecification. The inclusion of relevant covariates or better proxies for the size of locations may be required, or the use of higher levels of nesting, to ensure that equilibrium population scales proportionally, when considering aggregates of locations within nests.

3 Comparing with the existing literature

3.1 The traditional RUM framework for migration

Consider the gravity equation for migration as derived by Grogger and Hanson (2011) and many subsequent contributions (see Beine, Bertoli, and Fernández-Huertas Moraga, 2016, for an overview) based on the random utility maximisation framework of McFadden (1974). A potential migrant i in an origin country o compares utility among possible destination countries indexed by $d \in D$, among which the country of origin itself. Following the notation of Beine, Bertoli, and Fernández-Huertas Moraga (2016) as before, utility U of individual i is assumed to depend on an index of observables at the destination w_d, and on the bilateral

cost c_{od} of moving from o to d:

$$U_{odi} = w_d - c_{od} + \epsilon_{odi}.$$

With ϵ_{odi} EV distributed, the probability P_d of an individual in location o to prefer destination $d \in D$ over all other destinations $x \in D$ is

$$P_d = \frac{\exp(w_d - c_{od})}{\sum_{e \in D} \exp(w_e - c_{oe})}$$

If the number of prospective migrants is large such that $P_d = s_{od} = m_{od}/pop_{od}$ and defining $y_d = exp(w_d)$ and $\phi_{od} = exp(-c_{od})$ the following gravity equation is obtained:

$$m_{od} = pop_o y_d \phi_{od} \frac{1}{\sum_e y_e \phi_{oe}}.$$

A destination-mass variable may be included by the empirical researcher as a destination-specific explanatory variable in y_d, but its inclusion does not stringently follow from this popular theoretical framework. The corresponding log odds ratio is given by

$$\ln\left(\frac{s_{od}}{s_{oo}}\right) = \ln\left(\frac{m_{od}}{m_{oo}}\right) = w_d - w_o - (c_{od} - c_{oo}) \tag{6}$$

Again the researcher is left to decide whether the vector of explanatory variables should include some measure of size or dispersion, which would logically then be included in both in w_d and w_o.

3.2 Potential issues in existing empirical studies

In the framework of Section 2 where destinations are viewed as aggregates of choice alternatives, the number of alternatives in a location and the variation between them appear naturally as determinants of migration. In the traditional RUM based migration literature based on the framework described in the previous section, there is no choice within locations and no explicit role for the size of locations and dispersion within them. In empirical work often size and sometimes dispersion are added as explanatory variables in an ad-hoc fashion. However, given that these variables do not explicitly appear in the expressions, it is easy to make mistakes when doing so, and this section turns to considering some examples that may affected by this.

A source of confusion may be the fact that when considering the log-odds ratio as in

equation (5) in a framework where locations aggregate opportunities, the number of choice makers, pop_o drops out of the expression, but the number of alternatives in both origin and destination N_o and N_d remain. In an analysis based on equation (6) where size variables have to be added in an ad-hoc fashion, a researcher may wrongly believe that all mass variables cancel out and there is no need to include any measure of size, or perhaps that only the destination sizes N_d needs to be added.

Consider the seminal work of Grogger and Hanson (2011, p. 54) who include origin-destination (dyadic) fixed effects which capture size and other factors in an analysis based on the log-odds ratio similar to equation (6), omitting any size variable. In a secondary analysis, they consider the value of these estimated fixed effects as estimates of the 'fixed costs' of migration. Among all destinations considered they observe the largest residual attractiveness (as captured by the fixed effects) for the USA and Germany. Offered explanations are higher wages in these countries, labour-recruitment strategies in the 1960s, post-war asylum policies and immigrant networks. Whereas such factors may play a role, a more basic explanation for the large residual migration flows to these countries is that the USA and Germany are the largest destination countries in the dataset, and their analysis does not control for size.

Ortega and Peri (2013) and Beine, Bourgeon, and Bricongne (2019), estimate a dynamic version of the log odds equation (6) which may be stylised as

$$\ln(m_{odt}) = \ln(m_{oot}) + w_{dt} - w_{ot} - c_{odt} + \xi_{odt}.$$

Beine, Bourgeon, and Bricongne (2019) emphasise the importance of including origin-time fixed effects to control for m_{oot} and other origin-time-varying variables in w_{ot}. There is an asymmetry in their analysis, however, in that no destination-time fixed effects are included. Any time-varying measure of the number of opportunities in the destination or the dispersion between them present in w_{dt} would not be controlled for. In contrast to equation (6), in the suggested framework equation (5) shows explicitly that variables such as the number of jobs or dispersion in wages *in both origin and destination* determine migration flows and the log odds, and should be controlled for, suggesting that their analysis may suffer from omitted variable bias. Alternatively, destination-time fixed effects could have been included alongside the origin-time fixed effects to control for such factors.

In an innovative contribution Beine, Bierlaire, and Docquier (2021) consider a cross-nested logit model for migration. Their analysis assumes that the utility derived from choosing the origin does not depend on its size, whereas the utility of any other location does. The analysis in Section 2 rather suggests that the expected maximum utility from choosing any location increases with size, including for the origin region.

The relation between the coefficient on size and spatial phenomena described in Section 2.3 is easily misinterpreted. Estimating a gravity equation for migration in China from rural locations to cities Xing and Zhang (2017) find size coefficients close to 1.[10] They conclude that this explains the growth of larger cities. This seems unfounded since a coefficient of 1 on the size of the destination is compatible with a spatial equilibrium between spatial aggregates of different size without net migration, as argued in section 2.3. Urbanisation trends and agglomeration economies could rather be studied by including variables such as population density among the control variables, and a proxy for size or mass would be added to control for the more basic size differences occurring due to differences in the level of spatial aggregation of the observations, with an expected coefficient close to 1.

4 Empirical Application: Internal migration in Ethiopia

This section implements the framework of Section 2 where locations are viewed as aggregates of many choice alternatives, in an analysis of internal migration in Ethiopia. As richer specifications are considered the coefficient on destination size increases, as predicted. The predicted attractive effect of dispersion is confirmed, with a coefficient on the variance in the consumption per capita in the own region close to 0.5. The estimated coefficient is smaller for other locations, and smaller still for locations across ethno-linguistic borders. This runs counter to predictions of models using risk-aversion where dispersion in a location always is unattractive, and counter to the Borjas (1987) model where higher dispersion in the home region leads to more emigration. The fact that dispersion in outcomes is less attractive for farther destinations is to be expected if information is more difficult to obtain in such destinations such that uncertainty and risk aversion increasingly come into play. A last innovation is the inclusion of a weighted index combining several size variables.

4.1 Data Description

The main dataset used in the analysis is the 2013 wave of the Ethiopian labour force survey (LFS).[11] A recent study of internal Ethiopian migration using the LFS is Bundervoet (2018), who uses a multinomial framework and also considers qualitative aspects of migration. The LFS contains information on 240660 individuals. Such a large cross-section is important when studying migration which is a rare occurrence. We consider only individuals between 15 and 65 years old who have migrated in the 20 years prior to the interview or have never

[10] They find estimates below and above 1. Their preferred estimate is 1.056 with a standard error of 0.133

[11] The LFS can be downloaded freely from .

migrated, leaving 110615 individuals. About 9 percent of these report to have moved zone in the 20 years before the interview.

The number of variables in the LFS is limited, but crucially includes the current and previous zone of residence, and whether this (previous) place of residence is (was) in an urban or rural area within the zone. Migrants are also asked how many years ago they migrated.

We combine the LFS with data on housing from the Ethiopian Central Statistical Agency (CSA) 'Population and Housing Census of Ethiopia' from 2007,[12] and with the 2018 wave of the Living Standards Measurement Study (LSMS) for consumption expenditures[13]. We believe it is unproblematic to merge data from different years because the identification relies on cross-sectional variation. The relevant differences between zones driving our results, in terms of for example migration flows, housing stock, or population span several orders of magnitude and are persistent over time.

Although the LFS contains information on earned income, for several zones there are only a handful of sampled individuals with earned income, or none at all. This reflects the scarceness of paid jobs in these areas, which is taken into account in the analysis by controlling for the number of paid jobs as an independent variable. It is impossible to estimate the zonal mean or variance of earnings for zones without paid jobs, however. We therefore use the spatially adjusted consumption per adult equivalent from the LSMS to estimate the mean and dispersion in the return from opportunities at the zonal level, rather than earnings data. This variable is calculated with the explicit aim of measuring the standard of living of the individuals, including individuals who do not earn an income in monetary terms. Due to some border changes between zones that occurred between 2013 and 2018, combining the LFS and LSMS data implies that some small zones had to be merged.

There were 86 zones in Ethiopia in 2013. The LFS and both auxiliary datasets (on housing and consumption) differentiate between urban and rural areas within each zone. Some zones are purely rural or urban, however. We merge the 10 zones corresponding to the capital Addis Ababa. Others zones were merged due to border changes which are hard to trace: 4 small zones of the SNNPR region, the zones of the Gambela region and the zones of the Benishangul-Gumuz region. The Afar and Somalia regions are not considered because of the large share of semi-nomadic population. In total, the analysis considers 98 different locations. Appendix A provides a list of regions and zones included in the analysis, with some summary statistics, and an indicator for the zones which were merged. Considering

[12]This dataset is downloadable from the CSA website at and can be obtained in digitised form from the authors website or on request.

[13]This dataset is publicly available through the World Bank Central Microdata Catalog. See the project website for a description, technical documentation, and to download the microdata.

this many alternatives in a discrete choice model is computationally intensive. Often this is solved by considering the chosen alternative (migration destination), combined with a relatively small random sample from the set of non-chosen alternatives. We rather opted to keep the full set of alternatives and used extensive computing resources. Estimation was done using the maximum likelihood implementation of the Biogeme Python package (Bierlaire, 2020). This provides a convenient environment for handling data while allowing for the non-linear specifications of the utility functions which is required in the suggested framework when considering multiple size variables. All the datasets used in the analysis are publicly available. The Stata and Python code is available from the authors' website or on simple request.

4.2 Estimation equation and variable definitions

One of the richer specifications which will be brought to the data defines utility for an individual i from origin o from choosing destination d (allowing for $d = o$) as

$$\begin{aligned} U_{odi} &= \lambda \log(houses_d + b_j jobs_d) + \beta_c \log(cons_d) + \beta_v Var(cons_d) + \beta_u I(urban_d) \\ &\quad + I(o = d) \cdot (\beta_{oo} + \beta_a age_i + \beta_e educ_i + \beta_f I(female_i)) \\ &\quad + I(sameregion_{od}) \cdot \beta_s + \beta_d \log(distance_{od}) + e_{odi}, \end{aligned} \tag{7}$$

where e_{odi} is extreme value distributed. As in Ortega and Peri (2013) and Beine, Bourgeon, and Bricongne (2019), correlation in the error term e_{odi} is allowed for between destinations other than the origin, giving rise to a nested logit structure with the origin as a single alternative in a degenerate nest, and all other destinations grouped in a second nest. Write ξ for the dissimilarity parameter associated with this upper level of nesting. This basic structure captures and controls for the important fact that migrants are different from non-migrants in many ways that are hard to measure. A risk-averse individual may have a strong preference for the origin compared to any other destination, for example, which introduces correlation between destinations other than the origin.

The nesting of opportunities within each destination zone is only considered implicitly by the inclusion of a size variable for the destination. Specification (7) considers the weighted index $houses_d + b_j jobs_d$ as the size variable. The weight b_j will be estimated from data together with the other parameters. The coefficient λ on the size variable captures the dissimilarity between the opportunities proxied by the size variable, as discussed in the previous sections. It should not be confused with the dissimilarity parameter ξ which pertains to the dissimilarity between the origin zone and all destinations.

We variously consider population pop_d as a sole size variable, or the index $houses_d + b_j jobs_d$ combining the number of jobs $jobs_d$ (number of employment persons with earned income in the LFS) with the number of houses with running water $houses_d$. Some specifications will include the variance of the consumption in the destinations, $Var(cons_d)$, and an indicator $I(urban_d)$ for urban destinations. All specifications consider the average level of consumption in the destination zone $cons_d$, in logs.

If a constant would be added to the utility of every alternative it would not affect the probabilities and therefore would not be identified. The constant β_{oo} therefore only appears for the origin region, capturing all factors which make choosing the origin (i.e. not migrating) a more likely outcome. Similarly, variables such as the individual's age are modelled only to affect the probability of choosing to stay in the origin. Controls at the individual level include the age at the time of migration (we take the age at the time of the interview for non-migrants), a dummy for females, and the education level at the time of the interview. Education is measured on a 4-level scale which enters as a continuous variable to limit the number of parameters.

Origin-destination level controls include the distance $distance_{od}$ between the geographic centres of origin and destination zone, in logs, and an indicator whether the origin and destination zone are in the same region $I(sameregion_{od})$. The internal distance was taken to be 20km for all zones. Although this is a crude approximation, any error in scale will be captured by the own-region specific dummy.

Some specifications include interactions of variables with $I(o = d)$, for example to investigate whether the coefficient on the variance of consumption is different for the origin region versus when choosing a destination different from the origin. Likewise, interactions with $I(sameregion)$ will be considered.

4.3 Results

Table 1 presents the results. Column (1) considers a basic specification with population as the mass variable for the destination. The coefficient is less than 0.5, compared to the value of 1 expected in theory. A likely explanation is that the population size of a zone does not correlate strongly with the number of opportunities therein. Ethiopia is characterised by a large disparity in the level of development between localities: some populous rural low income zones offer few opportunities to migrants, whereas a city like Addis Ababa is both populous and offers many opportunities. The effect of distance is as expected. The coefficient on the dummy indicating a destination zone in the same region as the origin $I(sameregion)$ has the 'wrong' sign. This may be a further indication of a misspecified

model. The coefficients for the individual characteristics show the effect of these variables on the probability of not migrating. The effects are as expected: older or less educated individuals and woman are less likely to migrate. In Bundervoet (2018), in contrast, females are found to migrate more in Ethiopia.[14] The sign on gender will turn out to change between specifications. The very low value of the dissimilarity parameter ξ suggests that there is significant unobserved correlation between destinations other than the origin.

Column (2) introduces a dummy variable for the own region, capturing some of the unobserved part of utility that is specific to either the own-origin nest or the nest containing all other destinations (adding the dummy to the other nest would lead to the same result with the sign flipped). The dissimilarity parameter ξ for the upper level jumps from 0.155 to 0.242, suggesting that the simple dummy indeed captures some of this correlation.

Column (3) replaces the population in the destination with a weighted index of the number of houses and the number of jobs in the destination, as described in section **??**. The weight b_j of the jobs variable in the index is estimated together with the other model parameters. The coefficient on the combined mass variable is 0.77, compared to the coefficient of 0.472 when considering population as the mass variable. This value being much closer to 1 suggests that the index combining the number of houses and jobs is significantly better at capturing the size of the underlying opportunity-set in the destination. Intuitive properties such as scaling and aggregation then hold approximately and the model describes a situation closer to a spatial equilibrium, as described in section **??**. Also noteworthy is the change in the dissimilarity parameter ξ associated with the choice between the own region and any other region: this parameter further increases from 0.242 to 0.302, suggesting that relevant control variables have been added, reducing the correlation in the unobserved part of individual utility in the explicitly modelled nests, and bringing the model somewhat closer to the multinomial ideal. Moreover, the effect of per capita consumption drops significantly after introducing appropriate controls for the size of destinations, suggesting that this variable was partially capturing the effect of the abundance of opportunities in the destinations in the first two columns. Another sign that the specification with two mass variables in column (3) is to be preferred, is the fact that destination zones in another region now are estimated to be less attractive compared to those in the own region, as one would expect. This small effect of regional borders in partially explained by migration to Addis Ababa, the capital, which is highly attractive to migrants from all regions. In an unreported specification, adding a dummy for Addis Ababa to the specification of column (3) increases the effect of regional

[14] This may be related to the fact that individuals reporting are considered to have migrated from the same origin-zone as their current zone of residence (and also do not switch between rural or urban areas within the zone) as non-migrants, whereas Bundervoet (2018) also considers these intra-zone movements as migration.

	(1)	(2)	(3)	(4)	(5)	(6)
log(pop)	0.48	0.472				
	(0.011)	(0.011)				
log(houses + $b_j$jobs)			0.767	0.784	0.775	0.789
			(0.0065)	(0.00752)	(0.00768)	(0.00769)
b_j			0.479	1.78	1.49	1.614
			(0.0248)	(0.17)	(0.14)	(0.16)
log(distance)	-1.72	-1.7	-1.61	-1.59	-1.59	-1.6
	(.0103)	(0.0104)	(0.00951)	(0.0093)	(0.00926)	(0.00931)
log(cons)	2.07	2.06	0.838	0.293	0.31	0.274
	(0.0246)	(0.0249)	(0.0195)	(0.0223)	(0.0223)	(0.022)
I(urban)				1.13	1.05	1.05
				(0.0289)	(0.0288)	(0.03667)
Var(cons)						0.104
						(0.00396)
I(same region)	-0.499	-0.456	0.0566	0.0552	-0.0461	0.354
	(0.0239)	(0.0241)	(0.0231)	(0.023)	(0.0227)	(0.0479)
I(same region)·I(urban)					-0.136	-0.18
					(0.0431)	(0.0434)
I(same region)·Var(cons)						0.174
						(0.00882)
I(o=d)		2.91	3.52	4.29	6.92	2.45
		(0.128)	(0.109)	(0.132)	(0.129)	(0.117)
I(o=d)·age	0.461	0.248	0.203	0.242	0.196	0.181
	(0.0212)	(0.0104)	(0.00552)	(0.00691)	(0.0063)	(0.00614)
I(o=d)·educ	-1.72	-1.77	-2.02	-2.46	-1.92	-1.8
	(0.101)	(0.0668)	(0.0474)	(0.059)	(0.0691)	(0.0669)
I(o=d)·I(female)	0.265	-0.241	-0.253	-0.306	-0.239	-0.222
	(0.0984)	(0.0662)	(0.056)	(0.0665)	(0.0513)	(0.0502)
I(o=d)·I(urban)					-0.393	-0.595
					(0.096)	(0.121)
I(o=d)·Var(cons)						0.247
						(0.0289)
ξ	0.155	0.242	0.287	0.242	0.3	0.323
	(0.00767)	(0.00936)	(0.00652)	(0.00587)	(0.00993)	(0.00978)
AIC	228336	227930	214915	213278	213208	212334
BIC	228413	228016	215011	213384	213333	212487
N				110615		

Table 1: *Parameter estimates of a nested logit model for internal migration in Ethiopia. Robust standard errors in parenthesis.*

borders from 0.0566 to 0.177.

Column (4) introduces a dummy for urban destinations. Urban destinations are found to be more attractive. However, the lower dissimilarity parameter suggests that residual correlation has been introduced within the nests. We therefore allow the effect of the urban dummy to differ for the origin and for destinations in the same region (this includes the origin zone) in column (5). This substantially increases the estimated dissimilarity parameter, suggesting a better fit. Individuals are also more likely to choose their own region (not to migrate) if it is urban, with the effect of an urban origin on the probability of staying equal to 1.05-0.393-0.136=0.521. Urban zones within the same region are also more likely to be chosen, with an estimated effect of 1.05-0.136=0.914. For zones in a different region, the effect is largest at 1.05. Put differently, migration is estimated to be more likely from rural origins and to urban destinations. However, the attraction of a city is weakest for the origin region (deterring migration), stronger for cities in the same region, and strongest for cities outside of the own region.

Column (6), lastly, introduces the variance in annual consumption per adult equivalent at the zonal level as an additional explanatory variable. Also here differences in the effect are allowed between the zone of origin, zones within the same region, and zones in other regions. The attractive effect of dispersion in opportunities is found to be largest for the own region (0.104+0.174+0.247=0.525). It is smaller for other destinations in the own region (0.104+0.174=0.278), and smallest for destinations in other regions (0.104). These differences are statistically significant. It is reasonable to assume that information is more readily available on the availability and properties of opportunities in the own current location, or locations nearby (in the same region). The differences found in the attractive effect of dispersion then are in line with the model, which assumes that dispersion in the return to opportunities is attractive to individuals if they can observe and choose among the opportunities. Lastly, the estimated coefficient of 0.525 on the variance of consumption in the own region is very close to the predicted value of 0.5 in section **??**, equation (**??**), when assuming a true value for the dissimilarity parameter of 1.

A final observation is that the introduction of dispersion reduces the effect of the urban dummies. Also here, this reduction is strongest for the origin (deterrence of migration), less strong for other zones in the same region, and quite small for the destination zones in other regions. This suggests that the lack in local dispersion of opportunities may explain migration from rural areas.

5 Summary and Conclusion

This paper presented a random utility framework for migration where destination countries or regions are considered as aggregates of many alternatives between which migrants can freely choose. Migrants then consider the expected maximum utility when choosing a location, which increases with the number of alternatives and the dispersion between them in a location. The model serves as an extension or alternative to the prevalent specifications considering countries or regions as the fundamental unit of choice of migrants, where size and dispersion have to be considered in an ad-hoc fashion, without guidance from theory.

It was argued that the coefficient associated with size should be close to 1 in a well-specified model. Only then intuitive spatial properties hold such as independence of predicted flows or equilibrium population distribution from the level of spatial aggregation.

associated coefficient smaller than one, attenuating the effect of size. The traditional gravity equation where countries are the relevant unit of choice for migrants is obtained as a limiting case with perfectly correlated opportunities. In this case only properties at the country level which are unrelated to size, such as climate, average wage, or the unemployment rate, explain migration flows.

We showed that a coefficient on size equal to 1 is a property of an ideally specified model containing all relevant covariables and additional levels of nesting of locations. A coefficient on size substantially smaller than 1 (or omitting size altogether), is symptomatic of misspecification, and results in undesirable spatial properties: migration and the equilibrium distribution of population then depend on the level of spatial aggregation.

This result assumes that migrants can choose between opportunities at the destination, ignoring less favourable ones. This is only realistic if prospective migrants have sufficient information about the opportunities. In this case, destinations with equal average opportunities but more extremes opportunities are more attractive. The attractiveness of otherwise similar destinations with a wider variance in economic opportunities may be linked to trends of urbanisation in developing countries, where cities typically are characterised by very unequal economic outcomes; and with the observed overall attractiveness of destination countries with a more unequal income distribution in the context of international migration.

Practical implications for applied research are that (1) a size proxy for the destination should be included in gravity equations for migration. This proxy should be related to the number or mass of opportunities operating as an attractive force in the destinations. The associated coefficient reflects the dissimilarity between the underlying opportunities. A coefficient substantially smaller than 1 could point to a poorly defined model. (2) In log-odds expressions, the size variable capturing attractiveness through the number of available

opportunities in the destinations appears twice, in logs: once for the destination and once (with a negative sign) for the considered alternative (most often the location of origin). (3) If migrants are simultaneously looking for different types of opportunities (jobs, housing, etc.) the size variables are combined in a weighted index, the weights of which can be estimated from data. (4) If the utility from opportunities in a destination can be described stochastically, and migrants receive information on the specific opportunities and can choose between them, other things equal, dispersion of utility within a destination is an attractive factor and enters the utility function and gravity equation. For iid normally distributed opportunities, with a coefficient of 1 on the mass variable, the expected coefficient on the variance variable is 0.5.

The application to Ethiopian internal migration shows how the framework can be implemented and aims to further our understanding of the factors driving urbanisation. Two size variables were combined and the index weights were estimated from data. Dispersion in adult-equivalent consumption in destinations was considered, revealing a positive correlation with migration flows, as predicted. This effect is larger for the origin (discouraging migration), it is weaker for alternative destinations within the same region, and weakest for destinations outside of the own region. This is supportive of the hypothesis that the effect is stronger if more information is available. Controlling for dispersion in opportunities explains part of the attraction of urban origins. Put differently, the results suggest that lack of dispersion in opportunities in rural origins may be causing migration out of rural areas.

References

Anas, A. (1983): "Discrete choice theory, information theory and the multinomial logit and gravity models," *Transportation Research Part B: Methodological*, 17(1), 13–23.

Anderson, J. E., and E. V. Wincoop (2003): "Gravity with Gravitas: A Solution to the Border Puzzle," *The American Economic Review*, 93(1), 23.

Behrens, K., and Y. Murata (2021): "On quantitative spatial economic models," *Journal of Urban Economics*, 123, 103348.

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

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," *IZA Discussion Paper*, 14090.

Beine, M., P. Bourgeon, and J. Bricongne (2019): "Aggregate Fluctuations and International Migration," *The Scandinavian Journal of Economics*, 121(1), 117–152.

Berry, S. T. (1994): "Estimating Discrete-Choice Models of Product Differentiation," *The RAND Journal of Economics*, 25(2), 242–262.

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

Bertoli, S., J. F.-H. Moraga, and L. Guichard (2020): "Rational inattention and migration decisions," *Journal of International Economics*, 126, 103364.

Bierlaire, M. (2020): "A short introduction to PandasBiogeme," in *Technical report TRANSP-OR 200605. Transport and Mobility Laboratory, ENAC, EPFL.*

Borjas, G. J. (1987): "Self-Selection and the Earnings of Immigrants," *The American Economic Review*, 77(4), 531–553.

Bundervoet, T. (2018): "Internal Migration in Ethiopia, Evidence from a Quantitative and Qualitative Research Study," *World Bank, Washington, DC.*

Cardell, N. S. (1997): "Variance Components Structures for the Extreme-Value and Logistic Distributions with Application to Models of Heterogeneity," *Econometric Theory*, 13(2), 185–213.

Daly, A. (1982): "Estimating choice models containing attraction variables," *Transportation Research Part B: Methodological*, 16(1), 5–15.

Davies, R. B., and C. M. Guy (1987): "The Statistical Modeling of Flow Data When the Poisson Assumption Is Violated," *Geographical Analysis*, 19(4), 300–314.

Fally, T. (2015): "Structural gravity and fixed effects," *Journal of International Economics*, 97(1), 76–85.

Ferguson, M. R., and P. S. Kanaroglou (1997): "An empirical evaluation of the aggregated spatial choice model," *International regional science review*, 20(1-2), 53–75.

Fotheringham, A. S., and P. A. Williams (1983): "Further Discussion on the Poisson Interaction Model," *Geographical Analysis*, 15(4), 343–347.

Griffith, D. A., and M. M. Fischer (2013): "Constrained variants of the gravity model and spatial dependence: model specification and estimation issues," *Journal of Geographical Systems*, 15(3), 291–317.

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

Harris, J. R., and M. P. Todaro (1970): "Migration, unemployment and development: a two-sector analysis," *The American economic review*, pp. 126–142.

Kanaroglou, P. S., and M. R. Ferguson (1996): "Discrete Spatial Choice Models for Aggregate Destinations," *Journal of Regional Science*, 36(2), 271–290.

Katz, E., and O. Stark (1986): "Labor migration and risk aversion in less developed countries," *Journal of labor Economics*, 4(1), 134–149.

Kline, P., and E. Moretti (2014): "People, Places, and Public Policy: Some Simple Welfare Economics of Local Economic Development Programs," *Annual Review of Economics*, 6(1), 629–662.

Lerman, S. R. (1975): "A disaggregate behavioral model of urban mobility decisions.," Ph.D. thesis, Massachusetts Institute of Technology.

McFadden, D. (1974): "The measurement of urban travel demand," *Journal of Public Economics*, 3(4), 303–328.

——— (1977): "Modelling the Choice of Residential Location," *Cowles Foundation Discussion Papers*, 477.

Ortega, F., and G. Peri (2013): "The Role of Income and Immigration Policies in Attracting International Migrants," *Migration Studies*, 1(1), 47–74.

Persyn, D., and W. Torfs (2016): "A gravity equation for commuting with an application to estimating regional border effects in Belgium," *Journal of Economic Geography*, 16(1), 155–175.

Train, K. (2002): *Discrete Choice Methods with Simulation*. Cambridge University Press.

Wilson, A. (1967): "A statistical theory of spatial distribution models," *Transportation Research*, 1(3), 253–269.

——— (1970): *Entropy in urban and regional modelling*. London: Pion.

——— (1971): "A family of spatial interaction models, and associated developments," *Environment and Planning*, 3, 32.

Xing, C., and J. Zhang (2017): "The preference for larger cities in China: Evidence from rural-urban migrants," *China Economic Review*, 43, 72–90.

Appendix

Appendix A — Included zones and summary statistics

Table 2 gives a list of zones included in the analysis, together with summary statistics of the main variables. The sample used includes only individuals in the LFS that have never migrated or less than 20 years ago and who are between between 15 and 65 years old currently or at the time of migration. 'obs.LFS' pertains to the number of observation in our final sample derived from the LFS. pop 15-65 is the population of the zone estimated using the LFS sampling weights. Jobs is the estimated population-level number of jobs with paid earnings. Houses is the number of houses with a tap within the house or compound. 'consum.' is the nominal annual level of consumption per adult equivalent, spatially adjusted for food prices.

'MERGED' in the column Zone indicates that the line corresponds to a collection of merged zones within the region. Merging these zones was necessary to merge the LFS data with the LSMS data. All of the zones in the sparsely populated regions of Gamela and Benishangul-Gumuz were merged. In the SNNPR region containing a very large number of small zones, the zones Burji, Konso, Derash and the Segen Peoples' zone were merged.

Table 2: *Zones included in the analysis, with summary statistics.*

Region	Zone	Rur./Urb.	obs.LFS	pop(15-65)	jobs	houses	consum.
Tigray	North Western	Rural	916	738003	8479	417	11605
Tigray	North Western	Urban	308	117236	18645	7162	14564
Tigray	Central	Rural	1634	1112806	24360	1197	9660
Tigray	Central	Urban	560	260882	40431	15531	20655
Tigray	Eastern	Rural	784	555801	29224	1436	10241
Tigray	Eastern	Urban	1193	214502	31622	12147	29136
Tigray	Southern	Rural	1310	994514	39122	1923	10102
Tigray	Southern	Urban	412	146198	19477	7482	29031
Tigray	Western	Rural	428	340463	3734	460	10650
Tigray	Western	Urban	194	76726	8925	3432	23187
Tigray	Mekele	Urban	1278	264919	65167	0	32341
Amhara	North Gonder	Rural	1959	2850946	45881	6349	8833
Amhara	North Gonder	Urban	1873	567302	90459	34388	23038
Amhara	South Gonder	Rural	1515	2153115	60910	3609	11916
Amhara	South Gonder	Urban	398	276101	44638	15784	15192
Amhara	North Wollo	Rural	1093	1459643	34185	4507	9974

Region	Zone	Rur./Urb.	obs.LFS	pop(15-65)	jobs	houses	consum.
Amhara	North Wollo	Urban	300	195177	20083	18039	22823
Amhara	South Wollo	Rural	1809	2371873	52294	7626	7832
Amhara	South Wollo	Urban	2787	414613	71568	37080	31647
Amhara	North Shewa Amhara	Rural	1236	1684467	22970	4215	12380
Amhara	North Shewa Amhara	Urban	1512	287253	37152	30733	31366
Amhara	East Gojam	Rural	1602	1956255	47724	3054	8400
Amhara	East Gojam	Urban	1490	252745	38544	24028	13754
Amhara	West Gojam	Rural	1524	2125181	49548	3656	14120
Amhara	West Gojam	Urban	517	299153	28787	18394	16549
Amhara	Wag Himra	Rural	345	516981	10028	1201	3341
Amhara	Wag Himra	Urban	56	21820	6931	170	10365
Amhara	Awi/Agew	Rural	686	969006	36269	1716	11709
Amhara	Awi/Agew	Urban	236	128303	25516	12771	17535
Amhara	Oromia	Rural	269	377667	5052	2802	9321
Amhara	Oromia	Urban	111	73148	17655	8349	25814
Amhara	Bahir Dar Special	Urban	1280	199973	67352	33255	37100
Oromia	West Wellega	Urban	226	155389	25020	5085	17108
Oromia	East Wellega	Rural	765	1581861	33609	1753	11102
Oromia	East Wellega	Urban	1292	173846	35718	10654	12697
Oromia	Ilubabor	Rural	639	1141336	14215	1718	12839
Oromia	Ilubabor	Urban	269	155223	31420	7607	19313
Oromia	Jimma	Rural	1530	2498684	38894	3769	10305
Oromia	Jimma	Urban	229	147357	26545	3951	29148
Oromia	West Shewa	Rural	1088	1851274	33318	2615	12157
Oromia	West Shewa	Urban	391	321917	62012	23030	14644
Oromia	North Shewa Oromia	Rural	815	1358272	73819	3691	12197
Oromia	East Shewa	Rural	555	1035219	74971	6765	12652
Oromia	East Shewa	Urban	1838	442626	89180	44856	24972
Oromia	Arsi	Rural	1388	2525899	91616	4372	13322
Oromia	Arsi	Urban	1612	392751	61509	31754	24410
Oromia	West Harerge	Rural	1038	2226038	47245	3432	14115
Oromia	West Harerge	Urban	280	222586	41052	12073	20917
Oromia	East Harerge	Rural	1623	2810431	26646	9714	15522
Oromia	East Harerge	Urban	353	280888	18255	9065	25100
Oromia	Bale	Rural	691	1256138	22855	3869	15477
Oromia	Bale	Urban	275	241488	29749	20009	20399
Oromia	Borena	Rural	495	1073541	14207	1599	11305
Oromia	South West Shewa	Rural	670	1199779	13247	2848	9832

Region	Zone	Rur./Urb.	obs.LFS	pop(15-65)	jobs	houses	consum.
Oromia	Guji	Rural	694	1707576	10118	1738	13614
Oromia	Guji	Urban	216	141631	24155	8059	22543
Oromia	Jimma special	Urban	1399	155720	38615	14542	23025
Oromia	West Arsi	Rural	986	1940371	26367	5128	7674
Oromia	West Arsi	Urban	1562	397638	48230	24449	13559
Oromia	Kelem Wellega	Rural	488	812633	17603	1333	13399
Oromia	Kelem Wellega	Urban	145	79948	7653	2304	17687
Oromia	Horo Guduru	Rural	279	487584	33867	2520	12566
Benish.-G.	MERGED	Rural	2767	788836	15285	499	12295
Benish.-G.	MERGED	Urban	2086	156318	27892	302	23353
SNNPR	Gurage	Rural	1078	1144072	16721	4042	22565
SNNPR	Gurage	Urban	350	220342	39479	11541	21887
SNNPR	Hadiya	Rural	1103	1223226	25444	3070	13706
SNNPR	Hadiya	Urban	1436	168083	28130	10421	42814
SNNPR	kembata tembaro	Rural	574	624118	12856	906	6491
SNNPR	kembata tembaro	Urban	293	122834	20496	4850	7909
SNNPR	Sidama	Rural	2566	3006280	31079	7749	11712
SNNPR	Sidama	Urban	377	254024	32371	9879	24916
SNNPR	Gedio	Rural	688	757318	7421	1504	10610
SNNPR	Wolayita	Rural	1318	1438751	15606	4291	12939
SNNPR	Wolayita	Urban	1621	272785	45363	11628	18237
SNNPR	South Omo	Rural	475	573842	6264	901	6100
SNNPR	South Omo	Urban	114	57976	12626	1987	35429
SNNPR	Keffa	Rural	706	880847	24985	883	7874
SNNPR	Keffa	Urban	180	98390	13333	1806	10066
SNNPR	Gamo Gofa	Rural	1311	1586130	21092	3177	13430
SNNPR	Gamo Gofa	Urban	1495	229939	36760	15153	11656
SNNPR	Bench Maji	Rural	570	609542	14211	1122	6059
SNNPR	Bench Maji	Urban	229	136934	24737	1719	13093
SNNPR	Dawro	Rural	429	551500	13830	541	7491
SNNPR	Dawro	Urban	65	48928	11483	169	29510
SNNPR	Konta	Rural	131	170581	1771	282	7112
SNNPR	Selti	Rural	634	631856	13698	2548	7852
SNNPR	Selti	Urban	120	90954	18882	2175	33856
SNNPR	Alaba	Rural	243	250763	7678	635	6409
SNNPR	MERGED	Rural	459	602806	12645	369	4921
SNNPR	MERGED	Urban	124	54735	10326	4900	11418
Gambela	MERGED	Rural	2124	248060	6156	350	8819

Region	Zone	Rur./Urb.	obs.LFS	pop(15-65)	jobs	houses	consum.
Gambela	MERGED	Urban	2358	102926	17199	155	18638
Harari	Hareri	Rural	1626	96766	1336	440	15796
Harari	Hareri	Urban	2379	114248	24826	15108	25086
Addis Ababa	Addis Ababa	Urban	19196	3105712	892649	871494	22848
Dire Dawa	Dire Dawa	Rural	1609	140032	4051	823	14615
Dire Dawa	Dire Dawa	Urban	2392	244119	48724	20123	23222