Migrants looking for opportunities - On destination size and spatial aggregation in the gravity equation for migration

Persyn, Damiaan

DARE, University of Göttingen

11 December 2021

Migrants looking for opportunities

On destination size and spatial aggregation in the gravity equation for migration

Damiaan Persyn*

DARE, University of Göttingen

December 11, 2021

Abstract

I consider a RUM model for migration where destination countries or regions are viewed as collections of 'opportunities' which are the fundamental units of choice for migrants. The best opportunity for a prospective migrant is more likely to be found in a destination that has many and diverse opportunities. Recent contributions in economics studying migration rather consider entire regions or countries as the fundamental, atomistic, units of choice. The key role of the size of destinations and the diversity within them is therefore often not fully recognised, which may lead to biased inference. I argue that the coefficient on size equals 1 in the ideal RUM model. This is also required for the gravity model for migration to have some intuitive properties: only then migration flows scale proportionally when aggregating destinations, and there is zero net migration between otherwise similar regions of different size. Models omitting size or using a coefficient on size different from 1 violate these properties. Imposing proportional scaling also has implications for how different sets of opportunities should be combined. The approach is showcased in a study of internal migration and urbanisation in Ethiopia.

1 Introduction

The random utility based gravity equation used in current economic research on migration (see for example Beine, Bertoli, and Fernández-Huertas Moraga, 2016, for an overview) does

*email: damiaan.persyn@uni-goettingen.de. I am greatly indebted to Liesbeth Colen for helpful discussions and insightful comments that have greatly improved the paper. I 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 for helpful comments. Any remaining errors are mine.

not naturally include a size variable for the destination country or region. While many studies correct for this by adding destination size or destination dummies, a significant number of studies does not, or inconsistently. In reality gross migration flows are highly correlated with the destination size. Not controlling for the size of the destination then leads to poor predictions of migration flows if destinations differ in size; or bias in the estimation of the effect on migration of variables that are correlated with size due to omitted variable bias.

In this paper, I use the nested logit model of McFadden (1977) to extend the traditional framework by considering destination countries or regions as collections (nests) of many underlying atomistic units of choice for migrants. Depending on the context, a unit of choice could be a job, a dwelling, a partner, a piece of arable land, etc.İ call these 'opportunities' following the wording of Docquier, Peri, and Ruyssen (2014). Considering destinations as aggregates rather than as atomistic units of choice provides key insights which are missing from the traditional analysis. I summarise five such insights and point out the contributions of this paper in the next paragraphs:

A first key result from considering a nested logit structure with a set of underlying choices within destinations is that, under appropriate assumptions, the number of opportunities appears as an attractive factor in the gravity equation for migration, serving as the size or mass variable for the destination. I point out several well known studies that may suffer from biases as a result from considering destinations as fundamental units of choice and not controlling adequately for the size of the opportunity set in the destinations as a result.

Second, the probability that the best opportunity is found in a destination, and the expected migration flow to this destination, increases not only with the number of opportunities in the destination but also with the dispersion between them. Considering otherwise similar destinations, and assuming that migrants can choose between opportunities and avoid less attractive outcomes, heterogeneity in the utility from opportunities makes a destination more attractive. This property follows directly from random utility maximisation when considering aggregates of fundamental units of choice as described in McFadden (1977), but seems little known or applied. It may offer an alternative explanation for the observed attractive force of cities on 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). I do not know any migration study acknowledging the possible attractiveness

of dispersion in outcomes within destinations for homogeneous migrants.[1] I investigate the effect of dispersion in the empirical application to internal migration in Ethiopia, and find the predicted positive effect. This finding goes against those of Stark (2006). The effect is especially strong for the current residence, which seems consistent with the condition that migrants have good information on opportunities, for dispersion between them to become an attractive feature.

Third, it follows directly from the theory of discrete choice and the nested logit model of McFadden (1977) that the coefficient associated with the size variable reflects the dissimilarity of the opportunities within the destinations (see also Kanaroglou and Ferguson, 1996). This insight is lost in applications controlling for destination size in an ad-hoc fashion. In an ideal model including the most relevant factors affecting migration decisions residual correlation between opportunities should be small, opportunities perceived as dissimilar, and the mass variable should therefore have an associated coefficient close to 1. A minor contribution of this paper is to simply to remind researchers of migration of this interpretation of the coefficient on size. The deviation from its ideal value of 1 can therefore also be seen as a basic measure of model fit. A coefficient significantly smaller than one could point to a mass variable that does not proxy well for the number or the type of opportunities migrants are looking for in a destination, or point to significant correlation between opportunities within destinations, suggesting that an important control variable is missing or an additional level of nesting of destinations should be considered.

Fourth, a coefficient of 1 on size leads to two desirable spatial properties: (1) a coefficient of 1 on size is required to make predicted migration flows independent from the level of spatial aggregation, such that the predicted migration flow to a country equals the predicted migration flows to its constituent regions and vice-versa. This point was already made by Daly (1982). (2) I add to this by showing that a coefficient of 1 is required to have zero net migration between similar regions of different size. It is intuitive that there is little net migration migration between regions or countries with similar opportunities, even if the number of these opportunities is very different. There is indeed little net migration between similar countries of different size like Belgium and Germany, or between a typical region in a country and the rest of the country considered as an aggregate. These intuitive properties

[1]The mechanism at work makes a destination with a high variance in opportunities more attractive for all individuals, and is quite different from the sorting and self-selection of heterogeneous agents described by Borjas (1987).

do not hold in analyses omitting a mass variable, or when the coefficient on mass differs significantly from 1. To my best knowledge these consequences of a coefficient on mass or size deviating from 1 are not known in the migration literature.

Fifth, again following Daly (1982), if migrants are looking for different types of opportunities (for example jobs and arable land), and if predicted migration flows are to scale proportionally when considering aggregates of destinations as described in the previous paragraph, then the sizes of the opportunity sets should be combined (for example in a weighted index), which then enters the utility function and the gravity equation as the sole mass variable for the destination with an associated coefficient equal to 1. The appropriate weights of the different proxies for size entering the index can be estimated from data. Also this consistent aggregation of multiple opportunity sets (i.e. mass or size variables) has never been considered in the context of migration. I implement the aggregation of two mass variables in the empirical application to internal migration in Ethiopia.

A last contribution of this paper is to link important contributions in transportation science to the recent work in economics on migration. Examples are the seminal constrained models of spatial interaction of Wilson (1967, 1970, 1971); the estimation of gravity equations using Poisson regression by Flowerdew and Aitkin (1982); the work of Daly (1982), Anas (1983) and Kanaroglou and Ferguson (1996) who study aggregation by zones in spatial discrete choice models; and the relationship between dummies and 'balancing constraints' or 'multilateral resistance terms' studied by Fotheringham and Williams (1983); Davies and Guy (1987) and Griffith and Fischer (2013).

I apply the described methodology estimating a nested logit model for interregional migration flows in Ethiopia. An upper nest distinguishes between the own region and all other possible destinations as in Ortega and Peri (2013) and Beine, Bourgeon, and Bricongne (2019). The lower nest is implicit and considers two sets of opportunities in each destination region: the number of houses with running water and the number of jobs with paid earnings. These mass variables are combined in a single index as described above. Given that some regions hardly have individuals with paid earnings – or none at all, I consider the variance in consumption per adult equivalent at the destination to test for the attractive force of dispersion in opportunities. This variance correlates with more inward migration, as predicted. Moreover I find that the effect is significantly larger the for origin-region, which is supportive of the hypothesis that this attraction of variability in opportunities requires information to be available as assumed in the discrete choice framework.

The remainder of this paper is organised as follows: Section 2 considers the RUM based gravity equation for migration that is popular in current economic research. Section 3 introduces a nested logit model with a countable set of opportunities in destinations, which leads to the appearance of the number of opportunities as a size variable for the destination in the gravity equation. Section 4 considers extensions: dispersion in opportunities as an attractive force; multiple size variables per destination; and the link between the coefficient on the size variable and both aggregation properties and spatial equilibria. Section 5 considers the application to internal migration in Ethiopia. Section 6 concludes.

2 The traditional approach

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), which is based on random utility maximisation (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), 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:

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

Assuming that the error term ϵ_{odi} has an iid extreme value distribution results in a convenient expression for the probability P_{od} of an individual in location o to prefer destination $d \in D$ over all other destinations $x \in D$. Also using e as an index over destinations, it obtains that

$$P(U_{odi} > U_{oxi}) = \frac{\exp(w_d - c_{od})}{\sum_{e \in D} \exp(w_e - c_{oe})} \qquad \forall \ x \in D.$$

If the number of individuals is large, this probability corresponds to the share s_{od} of the population migrating from o to d. Writing pop_o for the number of individuals in o and m_{od}

for the number of migrants between o and d:

$$s_{od} \equiv \frac{m_{od}}{pop_o} = \frac{\exp(w_d - c_{od})}{\sum_{e \in D} \exp(w_e - c_{oe})}. \tag{1}$$

By bringing pop_o to the right hand side and writing y_d for $exp(w_d)$ and $\phi_{od} = exp(-c_{od})$ the resulting expression for the expected number of migrants m_{od} resembles a gravity equation:

$$m_{od} = pop_o y_d \phi_{od} \frac{1}{\sum_e y_e \phi_{oe}}. \tag{2}$$

The population in the origin appears naturally by assuming that there is a given number of individual decision makers in the origin. This number of choice-makers serves as the measure of the mass of the origin. It is perhaps disconcerting that there is no corresponding variable for the size of the destination, as you would expect in a gravity equation. Such a destination-size variable may or may not be included by the empirical researcher as a destination-specific explanatory variable in y_d, but its inclusion does not naturally follow from this traditionally used theoretical framework. Frequently the size of the destination is omitted.

After adding an appropriately defined multiplicative error term to equation (2) it can be estimated using Poisson pseudo-maximum likelihood. An advantage of the Poisson regression framework is that it is able to handle zero flows and heteroskedasticity, as emphasised by Silva and Tenreyro (2006). Also Flowerdew and Aitkin (1982) discussed the advantages of Poisson regression for the estimation of gravity equations.

An alternative frequently used in the literature is to consider the log of the ratio of two migration shares, the log-odds. A convenient choice is to divide by the share of stayers (individuals in o who prefer their current country of residence o over all possible destinations)

$$s_{oo} = \frac{m_{oo}}{pop_o} = \frac{\exp(w_o - c_{oo})}{\sum_{d \in D} \exp(w_d - c_{od})}. \tag{3}$$

Taking the log of the ratio of the shares (1) and (3) results in a linear function of the difference in the variables determining the attractiveness of a location.

$$\ln\left(\frac{s_{od}}{s_{oo}}\right) = \ln\left(\frac{m_{od}}{m_{oo}}\right) = w_d - w_o - (c_{od} - c_{oo}) \tag{4}$$

Again the researcher is left to decide whether the vector of explanatory variables should include some measure of size, which would logically then be included in both in w_d and w_o. This equation can be estimated by OLS. Advantages of this functional form are that it avoids numerical optimization and allows for instrumental variable estimation (Berry, 1994).

The omission of a mass variable for the destination can be problematic in empirical applications. Consider an analysis based on equation (1) or (4) for migration between two countries of vastly different size but with similar wages. An example could be Malta and Italy. Consider the case where size is not controlled for, wages are equal to 1 in both countries, and assume $c_{oo} = 0$ and $c_{od} = 4.6$ for $o \neq d$. For these values the model predicts 1 percent of inhabitants in either country will migrate. Malta has a population of about 500.000, so 5.000 Maltese would be predicted to migrate to Italy. Italy has a population of about 60 million, so 600.000 Italians would be predicted to migrate to Malta, more than doubling its population. This is obviously not realistic. The model does not reflect our intuition that migration flows also depend on the size of the destination. In real data the share of individuals choosing a destination likely depends on its size (as I will argue formally in the next section). An analysis on real-world data ignoring size is likely to find unexpectedly large residual flows to large destinations.

Such a pattern may be reflected in the results of Grogger and Hanson (2011, p. 54) who omit measures for the size of the destination in their analysis of international migration but control for size (and any other origin-destination specific factor) by including origin-destination (dyadic) fixed effects. In a secondary analysis, they consider the value of the estimated fixed effects from their main analysis as an estimate of '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 indeed play a role, a more basic explanation for the large residual migration flows to these countries is that these are the largest destination countries in an analysis that does not control for size.

Ortega and Peri (2013) and Beine, Bourgeon, and Bricongne (2019), estimate a dynamic version of (4) 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}. These origin-time fixed effects would also capture the effect of any time-varying size variable in w_{ot}. But there is an asymmetry in that no destination-time fixed effects are included, such that any time-varying effect of the size of the destination in w_{dt} (such as the number of jobs or the GDP which may be important given the focus on the business cycle in their analysis) would not be controlled for.

3 Putting back size in the RUM based gravity equation for migration

The previous section showed that the RUM based gravity equation for migration commonly used in the economics literature leads to a gravity equation which does not naturally include a size variable for the destination, and that this may be problematic in applied work if a size variable is not explicitly added (or if insufficient destination dummies are included; or if dummies are included but subsequently analysed without controlling for size).

As already noted by Kanaroglou and Ferguson (1996), the root of this problem lies with considering spatial aggregates as fundamental units of choice. This section introduces a countable number of fundamental units of choice in each destination, which we call opportunities as in Docquier, Peri, and Ruyssen (2014). I formalise the concept here drawing from a large and old literature on discrete choice in transport and spatial modelling (see for example McFadden, 1977; Flowerdew and Aitkin, 1982; Daly, 1982; Anas, 1983; Kanaroglou and Ferguson, 1996).

In a discrete choice setting, the probability that the maximum utility can be found in a destination (and therefore expected migration flow to this destination) increases with the number of fundamental units of choice found in the destination. If the utility derived from the opportunities within a destination is heterogeneous, the probability that the best opportunity can be found in a destination –and the expected migration flow to this destination– also increases with the degree of dispersion in opportunities. I consider this property as an extension in section 4.1.

3.1 Aggregating opportunities

Assume now that migrants do not seek countries but rather some type of opportunity, for example a dwelling or a job, contained in them. The opportunities can be contrasted with other determinants of migration that are frequently a such as climate, the average wage or a common colonial history. Such variables are clearly important to migrants and should be controlled for, but we cannot assign a number, size or mass to them.

Following Anas (1983) I first consider the case where utility is not correlated between opportunities. In this case, the equations (2) and (3) remain valid, with the adjustment that the choice now is no longer over destination countries, but over opportunities, which are indexed by $f \in F_d$ (for felicity), with F_d the set of opportunities in country d with cardinality N_d, such that the share of individuals in origin o choosing opportunity f in destination d equals

$$s_{ofd} \equiv \frac{m_{ofd}}{pop_o} = \frac{\exp(w_{fd} - c_{ofd})}{\sum_{e \in D} \sum_{g \in F_e} \exp(w_{ge} - c_{oge})}$$

When interested in the number of migrants to countries. one simply has to add the probabilities or shares corresponding to the opportunities contained within each destination country. Consider the case where the utility derived from the different opportunities within a country is identical such that $w_{fd} = w_d$ and $c_{ofd} = c_{od}$, then

$$\begin{aligned} s_{od} &= \sum_{f \in F_d} \frac{m_{ofd}}{pop_o} = \sum_{f \in F_d} \frac{\exp(w_{fd} - c_{ofd})}{\sum_{e \in D} \sum_{g \in F_e} \exp(w_{ge} - c_{oge})} \\ &= \frac{N_d \exp(w_d - c_{od})}{\sum_{e \in D} N_e \exp(w_e - c_{oe})}. \end{aligned}$$

Written as a gravity equation and writing y_d for $exp(w_d)$ and $\phi_{od} = exp(-c_{od})$ as before, we obtain

$$m_{od} = pop_o N_d y_d \phi_{od} \frac{1}{\sum_{e \in D} N_e y_e \phi_{oe}} \tag{5}$$

Considering countable, independent, atomistic units of choice (opportunities) within destinations, leads the size of the opportunity set in the destination N_d to appear naturally in the gravity equation.

The log-odds are

$$\ln\left(\frac{s_{od}}{s_{oo}}\right) = \ln\left(\frac{m_{od}}{m_{oo}}\right) = \ln(N_d) - \ln(N_o) + w_d - w_o - (c_{od} - c_{oo}).$$

Whereas the number of potential migrants in the origin pop_o cancels out when considering log-odds, the log of a size variable proxying the number of opportunities should be included for both origin and destination. The expected coefficients on these size variables are 1 and -1 given the assumption of strictly uncorrelated opportunities. The gravity equation and log-odds expression distinguish between the number of choice makers (for example the population pop_o) which operates as a push factor, and the number of opportunities N_o and N_d which operate as pull factors. One could use the population in the origin pop_o in the log-odds expression rather than N_o only in a context where a large population in a region correlates with an abundance of attractive opportunities. This assumption would be violated in an application considering countries or regions with significant differences in the level of development.

The gravity equation describing aggregate migration flows in equation (5) is asymmetric in that the size of the destination appears in the numerator and denominator, while the population as the size variable of the origin is absent from the denominator. Appendix B comments how this relates to an asymmetry in assumptions: whereas the number of choice-makers is constrained, the total inflow in each destination is not constrained. If one would also require predicted flows to match the observed aggregate inflows to each destination, one would obtain a fully symmetric gravity equation quite similar to that derived by Anderson and Wincoop (2003) in a CES framework. Their gravity equation in turn has the same form as the seminal doubly constrained model derived by Wilson (1970) from information theory some decades earlier, which is rarely acknowledged in economic studies.

3.2 Correlated opportunities: nested logit

It may be unrealistic to assume that the unobserved part in the utility an individual derives from different opportunities within the same destination is uncorrelated between them. Following McFadden (1977),I now consider a more general utility function with correlated opportunities (see also Kanaroglou and Ferguson, 1996; Train, 2002). Without loss of generality, the observed part of utility derived from choosing a certain opportunity f within a

destination (country, region,...) d is decomposed in a part w_d that is common among opportunities within d, and a part z_{fd} specific to the opportunity. Following the approach of Cardell (1997) and Berry (1994) which was first applied to migration by 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

$$U_{ofi} = w_d + z_{fd} - c_{od} + (1 - \lambda_d)\mu_{di} + \lambda_d \epsilon_{ofi}. \tag{6}$$

The unobserved part of utility which is shared among all opportunities within a destination for an individual, μ_{di}, is distributed 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 parameter λ_d or 'dissimilarity parameter' 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 of individuals between opportunities within a given destination.

We consider the following convenient decomposition of the corresponding probability that an alternative f within the opportunity set F_d of destination d is chosen:

$$\begin{aligned}
P_{o,f} &= P_{f|F_d} \cdot P_{o,F_d} \\
P_{o,F_d} &= \frac{\exp(w_d - c_{od} + \lambda_d I_d)}{\sum_e \exp(w_e - c_{o,e} + \lambda_e I_e)} \\
P_{f|F_d} &= \frac{\exp(z_{fd}/\lambda_d)}{\sum_{g \in F_d} \exp(z_{gd}/\lambda_d)} \\
I_d &= \log \sum_{g \in F_d} \exp(z_{gd}/\lambda_d).
\end{aligned}$$

If interest lies with predicting migration flows to destinations d (countries, regions, cities,...) which nest opportunities, rather than which opportunity is chosen within them, then only the aggregate level flow described by P_{o,F_d} and the log-sum or inclusive value term I_d are relevant. Under the assumption that the deterministic part of opportunity-specific utility is

constant within destination countries ($z_{fd} = z_d$), it holds that $I_d = \log(N_d) + z_d/\lambda_d$ and

$$P_{o,F_d} = \frac{m_{od}}{pop_o} = \frac{\exp(w_d + z_d + \lambda_d \log N_d)}{\sum_e \exp(w_e + z_e + \lambda_d \log N_e)} \tag{8}$$

or writing y_d for $exp(w_d)$ and $q_d = exp(z_d)$ and $\phi_{od} = exp(-c_{od})$ as before, it obtains that

$$m_{od} = pop_o y_d q_d N_d^{\lambda_d} \phi_{od} \frac{1}{\sum_e y_e q_e N_e^{\lambda_d} \phi_{od}}. \tag{9}$$

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, houses, arable land area, etc.), and the associated parameters $0 \leq \lambda_d \leq 1$ reflects how independent the unobserved part of utility is between opportunities in each destination. The log-odds at the country level then are given by

$$\ln\left(\frac{s_{od}}{s_{oo}}\right) = \ln\left(\frac{m_{od}}{m_{oo}}\right) = \lambda_d \ln(N_d) - \lambda_d \ln(N_o) + w_d - w_o + z_d - z_o - (c_{od} - c_{oo}).$$

One may obviously choose to test or impose the assumption that the dissimilarity between opportunities is equal among destinations ($\lambda_d = \lambda$).

If the correlation in the unobserved part of utility within destinations is large, opportunities within destinations are perceived as similar and the associated dissimilarity parameters λ_d will be small. If the model includes the most relevant control variables of locations and individuals, and considers other levels of nesting (which group destination countries or regions with similar properties such as language or ethnicity), then the residual unobserved component in utility (6) will be small, and opportunities will be perceived as dissimilar. In the limit, in a perfectly specified model, the parameter associated with the size variables should tend to 1.

A large opportunity set N_o in the origin makes it more likely that an individual will make the choice to stay in current location. This is clear in the utility $w_o + z_o + \lambda_o \log N_o$ obtained when choosing the origin (see (8)). The gravity equation (9) also shows how a larger N_o increases m_{oo} flows (the number of stayers) and lowers all other migration flows. It is especially clear in the log-odds expression where N_o appears with a negative sign for all destinations $d \neq o$. The discrete choice framework already considers the number of choice-makers in the origin, and this number of choice-makers appears as the mass variable in the

gravity equation. It is nevertheless also needed to include the number of opportunities N_o in utility functions or gravity equations when considering the choice to stay in the origin. This is not always done in practice. An example of an analysis omitting N_o may be Beine, Bierlaire, and Docquier (2021), who include population as a size variable affecting utility in all destinations, but do not include population for the origin.

These results suggest that an applied researcher — even if uninterested in the concept of underlying opportunities or the mathematics of nested logit models — would be well advised to include some measure of size of the destination in multiplicative gravity equations, or the log of the size of both origin and destination in log-odds regressions. The size variable should proxy the size of the opportunity sets in all locations, and the origin region itself is one such location. The associated coefficients serve as an inverse indicator of correlation between opportunities within destinations. A coefficient significantly smaller than 1 may suggest that the size proxy is not appropriate, that relevant control variables are missing, or further nesting of destinations should be modelled to reduce unobserved correlation within destinations.

4 Extensions

The previous sections argued that a proxy for the size of the opportunity set in the destination should be included in the gravity equation for migration, and showed how the coefficient on size relates to the perceived similarity of the opportunities to migrants. The framework with opportunity sets allows to consider other issues as well. In this section I first consider the case where opportunities are many and can be treated stochastically. The dispersion in the utility derived from opportunities can be shown to be an attractive factor of a destination. Second, I consider the case of migrants looking for different types of opportunities, and the nonlinearity this introduces in the gravity equation. Lastly, I revisit the coefficient on the size variable, and its link with spatial equilibria and the scaling or aggregating of migration flows.

4.1 Heterogeneous opportunities within destinations

If the observed part of utility associated with the opportunities within each destination can be described stochastically and is approximately iid normally distributed with mean z_d and

variance ω_d, McFadden (1977) shows that the utility of choosing destination d equals[2]

$$P_{F_d} = \frac{\exp(w_d + z_d + \frac{\omega_d}{2\lambda} + \lambda \log(N_d))}{\sum_e \exp(w_e + z_e + \frac{\omega_e}{2\lambda} + \lambda \log(N_e))}. \tag{10}$$

Underlying variability in the utility derived from opportunities makes a destination more attractive. If two destinations offer the same average return per opportunity, the maximum utility (and therefore the chosen opportunity) is more likely to be found in a destination with more opportunities and one with more dispersion in utility. This assumes that a migrant is able to observe and choose the opportunity within the destination. This result may be only relevant in situations with good information flows. A relevant could be migration over small distances, between regions of a developed country, where migration often occurs only after a jobmarket match has been made. Another example could be migrant networks passing on information on specific available jobs and other opportunities to prospective migrants in the home country. See also Bertoli, Moraga, and Guichard (2020) on how costs to information acquisition can shape migration decisions. Another case could be dispersion in opportunities found specifically in the current location of residence for which one can be expected to have quite good information.

The attractive effect of the variance in opportunities may offer an explanation for urbanisation trends in developing countries. Cities with a wide variation in opportunities are predicted to be attractive, in spite of high unemployment and other factors suggesting a low average expected return should a migrant arrive uninformed and unconnected. This explanation is different from the traditional analysis of Harris and Todaro (1970) where high wages in cities compensate for high unemployment rates, with prospective migrants considering expected urban wages taking into account the probability of becoming unemployed. The

[2]This section assumes $\lambda > 0$, McFadden (1977) considers the the limiting case $\lambda \to 0$ where the probability almost surely converges to

$$P_{F_d} \underset{\text{a.s.}}{\to} \frac{\exp(w_d + \max_f z_{fd})}{\sum_e \exp(w_e + \max_f z_{fe})}.$$

When opportunities are perceived as extremely similar conditional on w_d and z_{df}, their number becomes irrelevant and (apart from destination specific factors in w_d) only the maximum attainable z_{df} within each destination matters for the probability of destination d to be chosen. The properties z_{fd} of this best and only relevant opportunity within d can be absorbed in the destination specific variables contained in w_d. This case does not seem relevant for applications to migration.

discrete choice framework rather assumes that migrants ignore high urban unemployment rates because they can choose between opportunities, and focus on the higher wages offered by job opportunities in cities, possibly waiting until they obtain information on a specific opportunity presenting itself. This framework can also offer an explanation on why countries with high income dispersion are found be attractive in empirical studies on international migration.

4.2 Multiple sets of opportunities

As a second extension offered by considering sets of opportunities in destinations, imagine that migrants are heterogeneous and have different motivations for migration. In this case it may be unclear what size variable should be chosen, e.g. number of jobs, housing, landmass, etc. As argued by Daly (1982), rather than choosing between them, different size variables can be combined in a single weighted index. With different relevant size variables in the destination, $N_{1d}, N_{2d}, \ldots$ the utility from choosing destination d can be modelled as

$$w_d + z_d + \lambda \log(N_{1d} + b_2 N_{2d} + \ldots).$$

The same index $N_{1d} + b_2 N_{2d} + \ldots$ would enter the gravity equation as the combined mass or size variable for the destination. The index weights b can be estimated from data using maximum likelihood, although convergence can be more difficult. The log-odds are no longer linear in parameters and therefore no longer estimable by OLS.

Daly (1982) argues that the coefficient λ on the combined size-index should equal 1 if aggregating destinations is to lead to a proportional increase in predicted migration flows. The next section considers this argument more in detail.

4.3 Aggregation of migration flow, spatial equilibrium, and the mass variable in the gravity equation

Section 3.2 argued that in an ideally specified discrete choice model residual correlation in utility between opportunities within destinations should be small, and the parameter associated with destination size should be close to 1. This section considers neutrality to the level of aggregation and spatial equilibria as two more reasons why a coefficient close to 1 is a reasonable case to expect.

Consider the case of $R+1$ symmetric regions, indexed by r and k, with identical attributes $y_r q_r = 1$, populations $pop_r = pop$, and $N_r = N$ opportunities. Individuals face 0 migration costs when choosing their current region or $\phi_{rr} = exp(0) = 1$, and some positive cost when migrating such that $\phi_{rk}|_{r \neq k} = \phi$ (with $0 < \phi < 1$). The migration flow from r to k (see equation (9)) then is given by

$$m_{rk}|_{r \neq k} = pop N^{\lambda} \phi \frac{1}{N^{\lambda} + R N^{\lambda} \phi}$$

Now consider a scenario where R of the regions are considered as a single destination (say R regions are Italian) by the inhabitants of the remaining region $R+1$ (Malta). Consider the expected aggregate migration flow from Malta (region $R+1$) to Italy obtained from adding up the individual flows to each of the R Italian regions:

$$\sum_{r=1}^{R} m_{R+1,r} = R \cdot pop N^{\lambda} \phi \frac{1}{N^{\lambda} + R N^{\lambda} \phi} \tag{11}$$

When rather considering Italy as a single destination of size (rN), the predicted flow from Malta to Italy is

$$pop (rN)^{\lambda} \phi \frac{1}{N^{\lambda} + (rN)^{\lambda} \phi} \tag{12}$$

An obvious sufficient condition for these two perspectives in equations (11) and (12) to result in the same predicted number of migrants is $\lambda = 1$. This reflects the point made by Daly (1982) on requiring $\lambda = 1$ for migration to be proportional to the destination size, and for the gravity equation to be neutral to the level of aggregation of destinations. This perfect scaling with size fails if opportunities are seen as closer substitutes within locations compared to across locations. Although one may expect some deviation from perfect scaling, for most applications to migration scaling may be expected to hold to a large degree, and a coeffcient λ close to 1 on the mass variable would be reasonable to expect.

A further case for $\lambda = 1$ can be made by considering a spatial equilibrium. It is useful here to consider R as a measure of the size of the aggregated region relative to the remaining region. In the example Italy is R times larger than Malta since it contains R regions each of size N whereas Malta consists of only one such region. Continuing to assume that there are

migration costs between the R Italian locations, the total migration flow from Italy to Malta is given by the sum of the flows from the individual R Italian locations to location $R+1$, Malta. The expression for this aggregate migration flow from Italy to Malta corresponds to the right hand side of equation (11). A spatial equilibrium exists when this equals the flow in the opposite direction, from Malta to Italy, given by equation (12), or

$$R \cdot popN^{\lambda}\phi \frac{1}{N^{\lambda} + RN^{\lambda}\phi} = pop(RN)^{\lambda}\phi \frac{1}{N^{\lambda} + (RN)^{\lambda}\phi}$$

or equivalently

$$R \frac{1}{N^{\lambda} + RN^{\lambda}\phi} = R^{\lambda} \frac{1}{N^{\lambda} + (RN)^{\lambda}\phi}.$$

For $\lambda = 0$ the equation simplifies to $R = 1$: In case destinations are considered as atomistic as in the traditional analysis described in Section 2, migration flows are symmetric only if locations contain the same number of opportunities. If we assume that the number of opportunities in a location equals the population, this implies that the only stable spatial distribution of population is one with an equal population in both aggregate locations (Malta and Italy), however different their initial size.

For $\lambda = 1$, in contrast, the equation holds for any R: for any size difference R between two regions or aggregates of regions, the predicted migration flows in both directions will be equal, implying that any initial size difference is a stable spatial equilibrium. This is of course closer to reality: there tend to be only small net migration flows between countries or regions with similar properties (average wage, etc.), even if they are of very different size such as Malta and Italy.

The arguments in the previous section on the interpretation of the dissimilarity parameter together with those presented here on the conditions for scalability and spatial equilibria show that a reasonable value to expect for the mass variable in a well specified gravity equation for migration is close to 1. The link between the coefficient on size and spatial equilibria is also important when interpreting empirical estimates. Estimating a gravity equation for migration in China from rural locations to cities Xing and Zhang (2017) find size coefficients close to 1.[3] However, they conclude from this that 'migrants derive higher utilities from larger cities' and that this explains the growth of larger cities. This seems

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

unfounded as it was shown that in a basic gravity equation a coefficient of 1 on a size variable rather corresponds to the base case where migration flows do not alter a given spatial population distribution, i.e. a situation without any urbanisation trend. One can have a coefficient of 1 on size in combination with urbanisation if other variables are included, such as population density, or a dummy indicating urban destinations, and I will do so in our empirical application in the next section.

5 Empirical Application: Internal migration in Ethiopia

5.1 Data Description

The main dataset used in the analysis is the 2013 wave of the Ethiopian labour force survey (LFS).[4] A recent study of internal Ethiopian migration using the LFS is Bundervoet (2018), who uses the multinomial framework of section 3.1 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. I consider only individuals between 15 and 65 years old who have migrated in the 20 years prior to the interview or have never 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,[5] and with the 2018 wave of the Living Standards Measurement Study (LSMS) for consumption expenditures[6]. I 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

[4]The LFS can be downloaded freely from https://www.ilo.org/surveyLib/index.php/catalog/2363.

[5]This dataset is downloadable from the CSA website at https://www.statsethiopia.gov.et/census-2007-2/ and can be obtained in digitised form from the authors website or on request.

[6]This dataset is publicly available through the World Bank Central Microdata Catalog. See the project website https://microdata.worldbank.org/index.php/catalog/3823 for a description, technical documentation, and to download the microdata.

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. I 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. I 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 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. I 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.

5.2 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

$$\begin{aligned} U_{odi} &= \lambda \log(houses_d + b_j jobs_d) + \beta_c \log(cons_d) + \beta_v Var(cons_d) + \beta_u I(urban_d) \\ &\quad + I(o = d) \cdot (\beta_{oo} + \beta_a age_i + \beta_e educ_i + \beta_f I(female_i)) \\ &\quad + I(sameregion_{od}) \cdot \beta_s + \beta_d \log(distance_{od}) + e_{odi}, \end{aligned} \tag{13}$$

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 (13) 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.

5.3 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.[7] 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 4.2. 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 4.3. 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

[7] 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. I 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 4.1, equation (10), 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.

6 Summary and Conclusion

This paper presented a random utility framework for migration where destination countries or regions are considered to be nests of opportunities, and it are these opportunities which are the fundamental unit of choice of migrants rather than the countries or regions containing them. The model serves as an extension or alternative to the prevalent specifications considering countries or regions as the fundamental unit of choice of migrants, even if these destinations differ significantly in size.

If the opportunities are equally valuable to migrants and uncorrelated, their number appears as an attractive factor for the destination in the multiplicative gravity equation describing aggregate flows. If the unobserved part of utility is correlated among opportunities, the size variable in the aggregate gravity equation for migration has an 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 omitting the destination size, or using a coefficient on size substantially smaller than 1, leads to predictions that are violated in the data, such as large residual flows to large destinations in empirical studies, a counter-intuitive prediction of an equal spatial equilibrium distribution of population among locations that have similar characteristics up to size, or predicted migration flows that depend on the level of aggregation of the analysis. A coefficient significantly smaller than 1 therefore may rather point to a misspecified model, or a choice of a size variable for the destination that is a bad proxy for the number of attractive opportunities migrants are seeking.

If the deterministic part of utility derived from opportunities can be described stochastically, the variance in opportunities appears as an attractive factor of the destination. 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.

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.

Docquier, F., G. Peri, and I. Ruyssen (2014): "The Cross-country Determinants of Potential and Actual Migration," *International Migration Review*, 48(1_suppl), 37–99.

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

Flowerdew, R., and M. Aitkin (1982): "A Method of Fitting the Gravity Model Based on the Poisson Distribution," *Journal of Regional Science*, 22(2), 191–202.

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.

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.

Silva, J. M. C. S., and S. Tenreyro (2006): "The Log of Gravity," *Review of Economics and Statistics*, 88(4), 641–658.

Stark, O. (2006): "Inequality and migration: A behavioral link," *Economics Letters*, 91(1), 146–152.

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

Appendix B — Some remarks on the relation between discrete choice, constrained gravity equations and multilateral resistance terms

There is an asymmetry in how the size of the origin (the number of choice-making agents pop_o) and the size of the destination (the number of opportunities N_d) enter the gravity equation (5): whereas N_d appears in the nominator and denominator, pop_o only shows up in the nominator but not in the denominator. This stems from an asymmetry in the assumptions: The number of agents in each origin is given or fixed. Therefore if less individuals choose a specific destination, some other destination (or the origin) must experience a higher inflow from this origin. The number of arrivals in each destination is not fixed, in contrast. If less migrants choose a specific destination, there typically is no constraint enforcing that the decrease in inflows from one origin must be compensated by an inflow from another origin. Given that only the number of potential migrants in each origin is fixed, this model is known as the origin, production, or single constrained gravity equation (see Wilson, 1971). Taking the number of decision takers in each origin as given, but not the inflow per destination, seems particularly warranted in the study of 'supply driven' phenomena such as refugee flows or migration from underdeveloped countries where a decrease in the inflow from one origin to a specific destination does not imply an increase in the inflow from another origin.

The assumption of a fixed number of individual choice-makers in the origin is embedded in discrete choice frameworks such as multinomial logit and nested logit models. It can be implemented in a Poisson regression by including origin fixed effects, as emphasised by Fally (2015) in the context of international trade, but was already known by for example Fotheringham and Williams (1983); Davies and Guy (1987) and Griffith and Fischer (2013). The factor $1/\sum_{e \in D} N_e y_e \phi_{oe}$ in equation (5) assures that the constraint holds that $\sum_o m_{od} =$

pop_o. It is therefore called a 'balancing factor' by Wilson (1971), and corresponds to one of the 'multilateral resistance' terms of Anderson and Wincoop (2003) or Bertoli and Fernández-Huertas Moraga (2013).

In the context of regional migration, or migration between similar countries, it may be reasonable to also consider the total inflow in each destination as given. An example would be the case where opportunities are jobs that need to be filled. If an inflow from one origin to some destination decreases, the jobs in the destination will be filled by an increase in the flows from other origins. Strictly imposing this constraint gives rise to the Wilson doubly constrained model (see Wilson, 1971) which is isomorphic to the gravity model of Anderson and Wincoop (2003).[8] The doubly constrained model can be empirically implemented in a discrete choice framework by including destination specific constants. As the number of choice makers is fixed inherently in discrete choice models, the origin-constraint always holds[9]. Using a Poisson regression, the doubly constrained model is what is estimated when including origin and destination dummies (fixed effects). The estimated values of these dummies corresponds to the origin and destination 'balancing constraints' or 'multilateral resistance terms'.

This text focussed on the origin-constrained model which has received more attention in the recent economic literature on migration. However, although most studies derive the origin-constrained model from a discrete choice framework, many studies are subsequently – perhaps unknowingly – estimating a doubly-constrained model by including both origin and destination fixed effects in the Poisson regressions in their empirical implementation.

[8] Whereas Anderson and Wincoop (2003) derived their 'doubly constrained' model using CES preferences, Wilson (1970, 1971) used information theory (entropy maximisation) and Anas (1983) used discrete choice theory. See for example Persyn and Torfs (2016) for an application of a CES based doubly constrained model to commuting.

[9] Anas (1983) shows that the maximum likelihood estimate of the destination specific constant in a multinomial framework equals the expression for the balancing constraint (multilateral resistance term).