Using migration flows for non-market amenity valuation

Adam Theising and Dan Phaneuf[*]

March 15th, 2019

Contents

1 Introduction and motivation

Sorting models at the national scale have been used to estimate the economic value of nonmarket amenities such as air quality and climate. This class of models typically uses a set of Metropolitan Statistical Area (MSAs) as the objects of choice and explains households' residency in an MSA as a function of wages, housing rents, local amenities, and household characteristics. By observing tradeoffs between amenities and wages and rents (i.e. the net income for a household in a location) across the landscape, it is possible to infer the marginal willingness to pay for an amenity based on Tiebout logic.

In this paper we contribute to the sorting literature by examining the potential of a new data environment consisting of aggregate data on a large number of moves, for which we observe both origin and destination, and microdata providing details on a subset of the moves. We propose a generalization of the canonical approach that exploits the unique features of this data and addresses several methodological issues that are important for this class of models.

Most national sorting models have relied on a data environment in which households' current MSA and socioeconomic characteristics are observed. The vector of aggregate location shares across the landscape is interpreted as a spatial equilibrium resulting from utility maximizing location choices by the households. This spatial equilibrium forms the basis for a discrete choice estimation strategy whereby the linear-in-variables utility from a location depends on household net income, an average (location fixed) effect shared by all households, and a heterogeneity component defined by interactions between household and location characteristics. In most applications a subsequent linear regression is used to decompose average utility into an unobserved part and a part based on observable location attributes. The marginal willingness to pay for an attribute is estimated as the ratio of utility function parameters and the coefficient on income (the 'marginal utility of income').

We investigate this spatial equilibrium approach to valuation when migration flows are observed. To exploit our observation of both the origin and destination of moves, we propose a static model that begins with a moving household observed at its origin MSA. We assume a

[*]Dept. of Ag & Applied Economics, University of Wisconsin-Madison. Contact: theising@wisc.edu

conditional utility maximization framework whereby given its origin, a household selects the MSA that maximizes its utility. Using both aggregate- and household-level data environments, we find that the inclusion of migration flows in a spatial equilibrium model generates a gravity-type equation that identifies the marginal willingness to pay for an local amenities in a novel fashion.

In this preliminary draft of the paper, we focus on understanding the advantages and pitfalls of our modeling approach, and generating discussion thereon. To accomplish this we present an outline of the modeling framework in the next section, followed by a description of our data and intent to revisit national (US) valuations of air quality and climate with this fresh sorting approach.

2 Modeling framework

Our migration model builds on the canonical random utility approach commonly adopted in the context of locational choice. Let household i, originating in location k and moving to location j, receive utility

$$U_{ij}^{k} = U(C_{ij}, P_{ij}, X_j, MC_{ij}) + \varepsilon_{ij}^{k} \tag{1}$$

where C_{ij} is consumption of the numeraire good, P_{ij} are local housing expenditures, MC_{ij} are the household's moving costs to location j, X_j is a vector of the location's amenities, and ε_{ij} is a separable, individual-specific, and idiosyncratic component of utility. $U(\cdot)$ is a function that reflects standard neoclassical microeconomic behavior.[1] Putting this together in words, a household's utility is derived from the economic characteristics of their *chosen* location and the demographic characteristics of the household. In this static framework, we assume a household's observed location choice results from them maximizing their utility subject to their budget constraint

$$W_{ij} = C_{ij} + P_{ij} \tag{2}$$

with W_{ij} as the household's local income. Thus, we can write the household's indirect utility function for choosing location j as

$$V_{ij}^{k} = U(W_{ij} - P_{ij}, X_j, MC_{ij}) + \varepsilon_{ij}^{k}. \tag{3}$$

If we assume ε_{ij}^{k} is i.i.d. type I extreme value, McFadden (1984) shows the probability of household i settling in location j can be expressed as

$$\Pr[V_{ij}^{k} \geq V_{il}^{k}, \forall l \neq j] \equiv \Pr_{ij}^{k} = \frac{\exp\left(U(W_{ij} - P_{ij}, X_j, MC_{ij})\right)}{\sum_J \exp\left(U(W_{ij} - P_{ij}, X_j, MC_{ij})\right)} \tag{4}$$

which in turn yields the familiar odds ratio

$$\frac{\Pr_{ij}^{k}}{\Pr_{ik}^{k}} = \frac{\exp\left(U(W_{ij} - P_{ij}, X_j, MC_{ij})\right)}{\exp\left(U(W_{ik} - P_{ik}, X_k, 0)\right)}. \tag{5}$$

This mechanical relationship between the mover-stayer odds ratio and utility as a function of a location's characteristics will prove valuable for inferring the value of a location's amenities. We will return to it below after we add structure to the model by parameterizing household utility.

2.1 *Empirical model*

We now take the model to a data environment, beginning in a cross-sectional context. Suppose we have two types of migration data: a representative sample of M moves, where the origin

[1] Continuous, twice-differentiable, etc. $U_i(\cdot)$ is also possible here, and will be incorporated in future work.

and destination are observed for each move, and for a subset $I < M$ of the moves, detailed information on characteristics of the migrating households. We refer to these as the aggregate and micro-level data, respectively, and again use $i = 1, \ldots, I$ to denote elements (i.e. households) of the micro-level data.

Denote by $j = 1, \ldots, J$ the universe of migration locations under analysis. Consider moves that originate at origin $k \in J$, and let $J_k \subseteq J$ denote the destination alternatives that are in the choice set for origin k households, which allows for the possibility that not all locations in the landscape are considered. Let indirect utility for a household i starting in location k and then migrating to destination j be

$$V_{ij}^k = \delta_j^k + \mu_{ij}^k, \; j = 1, \ldots, J_k \tag{6}$$

where δ_j^k is the *average utility* of moving to location j among households starting in place k, and μ_{ij}^k is an household-specific component of utility.

On a first pass, we assume μ_{ij}^k is i.i.d. type I extreme value. For the sake of tractability, we linearly parameterize average utility to a given destination from a given origin as

$$\delta_j^k = \beta X_j + \gamma X_j^k + \alpha \bar{Y}_j^k \tag{7}$$

where X_j is a vector of (observed) destination-specific amenity characteristics, X_j^k are (observed) migration cost factors with dyadic variation (e.g. distance between locations, historic migrant network intensity, etc.), and $\bar{Y}_j^k$ is the average *net* income $(\bar{W}_j^k - \bar{P}_j^k)$ of households who move from k to j.

Given the error structure of (6) and the aggregated average nature of (7), we can write the parameterized equivalent of (5) as

$$\frac{\Pr_j^k}{\Pr_k^k} \equiv \frac{N_j^k}{N_k^k} = \frac{\exp\left(\beta X_j + \gamma X_j^k + \alpha \bar{Y}_j^k\right)}{\exp\left(\beta X_k + \alpha \bar{Y}_k^k\right)}. \tag{8}$$

where N_j^k denotes the aggregate flow of migrants from k to j and N_k^k is the total flow of individuals remaining in location k, as observed in our aggregate data. Manipulating this into a log-odds ratio yields

$$\ln\left(\frac{N_j^k}{N_k^k}\right) = \beta(X_j - X_k) + \alpha(\bar{Y}_j^k - \bar{Y}_k^k) + \gamma X_j^k. \tag{9}$$

Variations on this equation form the primary backbone of our aggregate-level analysis. As noted originally in Anas (1983) and more recently in the international migration literature by Beine et al (2011) and Grogger and Hanson (2011), (9) is a restricted gravity equation, and describes the relationship between bilateral migration flows and the relative attractiveness of locations. More relevant to our application, this relationship also provides a novel source of parameter identification for nonmarket local amenity valuation.

Recall that we are interested in measuring the marginal willingness to pay for a single amenity, X_g. Taking the total differential of indirect utility as parameterized in (6) and (7), we have

$$dV_j^k = dX_j\beta + dX_j^k\gamma + d\bar{Y}_j^k\alpha. \tag{10}$$

Setting all differentials to zero except for $X_{j,g}$ and the net income, $\bar{Y}$,

$$0 = dX_{j,g}\beta_g + d\bar{Y}_j^k\alpha \Rightarrow \frac{d\bar{Y}_j^k}{dX_{j,g}} = -\frac{\beta_g}{\alpha} \tag{11}$$

giving us the average marginal willingess to pay for changes in X_g.

Both parameters needed for this measure can be estimated non-linearly from (9), using the Poisson Pseudo Maximum Likelihood (PPML) procedure described in Santos Silva and Tenreyro (2006). Despite requiring only aggregate data on migration flows, average net income, and local amenities, these parameters are microfounded by the random utility locational choice model.

While we discuss using microdata to capture household-level preference heterogeneity in the next subsection, it's valuable to highlight that some demographic heterogeneity can be modeled using only aggregate data. While demographic attributes of households do not vary over alternatives in a discrete choice setting, it is possible to include then in the numerator of (9) through the normalization of its counterpart in the denominator. See Li et al (2011) for an example in the context of health, environment, and automobile demand.

2.2 *Incorporating microdata*

Recall that our data environment also includes household microdata on a subset of migrants. We return to (6), where we defined a household's indirect utility quite generally, and seek to make use of the additional demographic information available in the microdata.

To begin, we assume a tractable and separable structure for the components of household-specific utility:

$$\mu_{ij}^k = \alpha f(Y_{ij}^k) + \gamma f(Z_{ij}^k) + \eta_{ij}^k \tag{12}$$

where Y_{ij}^k is household-specific income net of housing costs for household i in location j[2], Z_{ij}^k is set of household-characteristic and location-characteristic interaction terms that vary across dimensions i and j (or i and k), and η_{ij}^k is an i.i.d. random variable distributed type 1 extreme value. Additionally, we assume a linear structure for the average utility term, δ_j^k:

$$\delta_j^k = \beta X_j + \gamma X_j^k + \xi_j^k \tag{13}$$

where X_j and X_j^k are still destination-specific and dyadic characteristics and ξ_j^k is an idiosyncratic error.

Again, the bifurcation of indirect utility into household-specific and average utiliies allows us to take advantage of the model's random utility structure. Borrowing from the demand-analysis literature, we see that our migration model is closely-related to the so-called micro-BLP model (Berry et al, 2004). As such, while estimating our discrete choice model by maximum likelihood, we can use the BLP contraction mapping routine to numerically invert out the average utility δ_j^k parameters. Conditional on household-specific heterogeneity estimated in the "outer" maximum likelihood loop, this "inner" contraction mapping ensures our model's logit-predicted migration flow share matches the share of movers originating in city k that select city j as observed in our data. Thus, with some abuse of notation, δ_j^k are J sets of J_k conditional migration shares from origin k.

These δ_j^k parameters can be used to highlight a now microdata-driven relationship between *conditional* migration shares and location amenities that is the household-level analogue to the aggregate gravity equation outlined in (9). Consider δ_j^k and δ_k^k and note that

$$\delta_j^k - \delta_k^k = \beta(X_j - X_k) + \gamma X_j^k + (\xi_j^k - \xi_k^k) \tag{14}$$

To identify average utilities (ASCs) in this discrete choice model, we must fix an arbitrary element of each δ^k vector to zero and normalize the remainder of δ_js in that vector. The origin component, δ_k^k, is a natural candidate, as k is by default a member of the choice set J_k. This normalization

[2]Note that destination-specific counterfactual measures of household net income must be imputed, since thisinformation is only available for a household's *actual* location. See Bayer et al (2009) for the hedonic procedures used in light of potential Roy sorting.

results in

$$\hat{\delta}_j^k = \beta(X_j - X_k) + \gamma X_j^k + (\tilde{\xi}_j^k - \tilde{\xi}_k^k) \quad (15)$$

which again provides a micro-founded second-stage equation that can be estimated to recover our first structural parameter of interest, β. The other parameter necessary to measure marginal willingness to pay, α, is estimated by the outer loop of the maximum-likelihood routine.

3 Data

Our application of these models aims to revisit the nonmarket valuation findings of Bayer et al (2009) and Sinha et al (2018). Both papers model their spatial equilibria nationally, with metropolitan statistical areas (MSAs) serving as the location choice. The former values air quality, an endogenous amenity, while the latter values climate, which has typically been considered exogenous. The model framework laid out in section 2 is well suited to value both exogenous and endogenous amenities, and we follow the aforementioned in setting MSAs as a household's object of locational choice. Our framework described a particular use for both aggregate- and micro-data. We describe them now.

We've acquired aggregate measures of MSA-to-MSA migration flows from two separate sources. The first comes from the United States Internal Revenue Service (IRS) who produces an annual county-to-county report of migration based on previous- and current-year tax filings. We aggregate these county measures up to MSAs[3], and find them particularly attractive in that there is year-to-year variation in migration rates.[4] The major downsides of this aggregate data are that the data's migration patterns are only reflective of households who file tax returns, and privacy concerns result in the IRS censoring migration flows of less than 10 households. The second source of aggregate-level data is the US Census Bureau, who also provides county-to-county reports of migration based on American Community Survey responses. Again, these county migration flows have been aggregated up to MSAs. This data is only available in rolling 5-year aggregations, however, providing slightly less temporal clarity on when migration occured.

Household-level microdata has been obtained from public-use microdata (PUMs) created by the Census Bureau. PUMs are a 1%, nationally-representative sample of US households, taken annually. A different vintage of this micro-data was used in both national sorting models described above, and the data's principle advantage is its provision of highly detailed information on household's demographics, employment, and location. We've collected annual samples of this microdata since 2012; crucially, in these most recent years, we've been able to match surveyed households that moved to both their current and previous MSAs.

Finally, we've collected a panel of MSA-level data for several amenities going back to 2010. Of most direct relevance to our nonmarket valuation application, we have assorted measures of annual air quality (AQI, $PM_{2.5}$) and climate (temperature min/max, precipitation). But we have also collected annual measures of a variety of additional MSA-level amenities that urban and environmental economists have shown to be important to quality-of-life.

References

[1] Anas, A., (1983). "Discrete choice theory, information theory, and the multinomial logi and gravity models." *Trans. Res. B*, 13-23.

[2] Bayer, P., Keohane, N. and C. Timmins, (2009). "Migration and hedonic valuation: the case of air quality." *JEEM*, 58: 1-14.

[3] Beine, M., Docquier, F. and C. Ozden, (2011). "Diasporas." *JDevE*, 95: 30-41.

[3] An MSA typically consists of anywhere from 1 to 10 counties that make up an urban and suburban core

[4] We intend to take advantage of this potential year-to-year panel structure in ongoing work.

[4] Berry, S., Levinsohn, J. and A. Pakes, (2004). "Differentiated products demand systems from a combination of micro and macro data: the new car market." *JPE*, 112(1): 68-105.

[5] Grogger, J. and G. Hanson, (2011). "Income maximization and the selection and sorting of international migration." *JDevE*, 95: 42-57.

[6] Li, S., Liu, Y. and J. Zhang, (2011). "Lose some, save some: Obesity, automobile demand and gasolien consumption." *JEEM*, 61: 52-66.

[7] McFadden, D. (1984). "Econometric analysis of qualitative response model." In Griliches, Z. and M. Intriligator (eds.), *Handbook of Econometrics, Vol. 2*, North-Holland, Amsterdam.

[8] Santos Silva, J.M.C. and S. Tenreyro, (2006). "The Log of Gravity." *ReStat*, 88(4): 641-658.

[9] Sinha, P, Caulkins, M. and M. Cropper, (2018). "Household locaiton decisions and the value of climate amenities." *JEEM*, 92: 608-637.