Choice-Based Estimation of Alonso’s Theory of Movement: Methods and Experiments

GUOXIANG DING and MORTON E. O’KELLY
Department of Geography, Ohio State University, 1036 Derby Hall, 154 North Oval Mall, Columbus, OH 43210, USA
Tel: 1-614-292-2704; 1-614-292-6031 Fax: 1-614-292-3656
E-mail: ding.45@osu.edu; okelly.1@osu.edu

Abstract. Alonso's theory of movement (ATM) provides a general framework for interaction in a spatial context. Though conceptually elegant, application of this model is limited due to the difficulty in estimating model parameters. In this paper, the ATM is applied to trade area delimitation using choice-based data in a novel way. The approach permits flexibility with regard to model assumptions about flow constraints; that is, the model is neither attraction-constrained nor production-constrained. Data in a block group scale and ZIP+4 scale have been examined to estimate model parameters by spatial choice comparisons and OLS. The data at both scales produce reasonable descriptions of an inflow elasticity variable (approximately 0.6) and spatial separation (distance-decay parameter is approximately 1 at ZIP+4 and 2 at block group level). Block group level aggregation is adopted for the prediction of conditional probability of spatial choices, since this level of aggregation gives a better fitted model (R-square is 0.59). This scale produces a more robust result as there are more paired comparisons in the block group level data, and thus the data incorporate more travel observations compared with ZIP+4 level data. The probability of customers' spatial choices is related with the store properties (attractiveness, relative accessibility), and the spatial separation between origins and destinations. Through empirical examination, we observe that store attraction increases at a decreasing rate, with respect to store size (measured by floor space).

1 Introduction

Spatial Interaction (SI) is defined as movement or communication over space that results from a decision process, in which decision makers trade off benefits from interaction with the costs needed to overcome spatial separation (Fischer 2000; Fotheringham and O'Kelly 1989). The family of SI models can be divided into four types: unconstrained, production-constrained, attraction-constrained and doubly-constrained, which are distinguished by origin and destination flow constraints embodied in them (Fotheringham and O'Kelly 1989). Getis (1991) argued that SI models formally have a broad range of equivalence to statistics of spatial autocorrelation and they can be used to capture embedded patterns in spatial processes and spatial structure (Roy and Thill 2004). For instance, SI models have been widely recognized as an important model applied in migration simulation, business and goods flows, travel-to-work flow allocation, capital movements, and recreational trips over geographic space (Fotheringham and O'Kelly 1989). SI models also have been proven to be effective in spatial choice modeling (Miller and O'Kelly 1990). Typically, production-constrained SI models have been employed to simulate the spatial choice process, which, by definition, assumes fixed flows emanating from each origin.

Alonso (1973; 1978) proposed a generalized framework in which four types of spatial interaction models mentioned above can be described as special cases. Specifically, quasi-constraint features have been incorporated into the framework, which may then be applied in circumstances where the marginal flow totals are uncertain (Fotheringham and Dignan 1984). This framework is called as Alonso's theory of movement (ATM) which can be treated as a general form of SI models [more detailed discussion on Alonso's theory of movement is in section 2]. ATM is very flexible in modeling socioeconomic movements and has achieved much attention in the study of migration, urban traffic, social mobility, international trade, etc. (Hua 2001). In addition, it is interpretable in decision making (Roy and Thill 2004).

Model calibration is an important consideration in the application of both SI models and ATM. Several important ideas, such as entropy maximization, information theory models, a cost efficiency principle and most probable states, spatial structure and spatial interdependencies, and neural networks have contributed to the conceptual development of these models (Roy and Thill 2004). Two kinds of sampling techniques, random sampling and choice-based sampling, have been adopted into model calibration so far. For example, SIMODEL, written in FORTRAN, employs maximum likelihood estimation to calibrate traditional SI models using random samples and has been proved to be robust and efficient (Williams and Fotheringham 1984). Miller and O'Kelly (1991) presented properties for a production-constrained ATM using simulated random samples, Ordinary Least Square (OLS) and Maximum likelihood (ML). They demonstrated that parameters in ATM can be recovered from random simulated data if exogenous variables are measured correctly. Random sampling designs are easy to understand and implement with consistent and effective parameter estimators (Thill and Horowitz 1991). However, it is often the case that a random sample produces relatively few respondents to a specific chosen alternative, and interviewing lots of people who have not chosen a particular mode or destination seems to be quite inefficient. In addition, it is also expensive and difficult to apply random sampling to acquire certain accurate information for analysis (Manski and Lerman 1977; Thill and Horowitz 1991). Random sampling also has strict requirement of homogeneity. All these restrictions limit the practical application of random sampling in model calibration. Choice-based sampling becomes one good alternative.

Choice-based sampling collects data on-site by interviewing decision makers who have made (or are in the course of making) actual spatial choices. For example, surveying transit users at the station and auto users at the parking lot provides a cheaper way to study travel behavior compared to surveying people at home. Similarly, when studying shopping destination choices, conducting survey at a store is much more convenient and efficient than home interviews (Manski and Lerman 1977). Choice-based data are cost-effective because the data collection is from clustered decision makers who are easily observed at their chosen alternatives (Manski and Lerman 1977; Waldman 2000). Such data may include valuable information on variables that determine existing choices, and the ultimate goal of choice-based data is to infer or calculate the behavior of population at large from the choice-based observations (O'Kelly 1999). Application of choice-based samples gives rise to a semiparametric inference problem and solutions to this difficult problem can be found in several literatures (Lancaster 1997). Cosslett (1981) developed a "pseudo-likelihood," a quantity closely related to the log likelihood for a random sample, to investigate the estimator in discrete-choice models using choice-based samples, and the estimator was proved to be consistent and asymptotically normally distributed. Several other estimators also were obtained and proved to be consistent for choice-based samples (Imbens 1992; Manski and Lerman 1977; Manski and McFadden 1981). Thill and Horowitz (1991) discussed techniques to estimate the model parameters in multinomial logit and other probabilistic discrete-choice models using choice-based samples with limited information, and they also discussed additional difficulties in choice-based estimation compared with using random samples. Following the methods by Cosslett (1981), they recovered and explained important information about spatial choices from choice-based data. O'Kelly (1999) applied choice-based data to calibrate a trade area gravity model, which allowed the efficient use of easily collected data to infer the distance-decay rates and store attraction parameters. One recently developed paper by O'Kelly (2005) employs choice data to estimate production-constrained SI model parameters for retail trade area using OLS with consideration of store competition and agglomeration. This is an improved and simpler method compared with (O'Kelly 1999) by data transformation. However, the application of choice-based data in the context of Alonso’s model remains untouched so far.

Though ATM is elegant and powerful, one prominent barrier to ATM’s wider application is the estimation difficulties and complexity due to the model inflow and outflow endogeneity. Discussions on ATM calibration generally apply regression (ordinary and weighted least squares) or maximum likelihood methods with separate considerations of systemic variables and distance-deterrence functions. A good deal of related empirical work has been reported in the past decade (Alonso 1973; Anselin 1982; De VOS and Bikker 1982; Hua 2001; Ledent 1980; Miller and O'Kelly 1991; Mueser 1989; Poot 1986; Porell and Hua 1984; Tabuchi 1984; Vries, Nijkamp *et al.* 2002). In this paper, we derive a spatial model from ATM and a method is proposed to calibrate ATM parameters using choice-based data. This work has two significant contributions. First, ATM framework is extended to simulate the spatial choice in a probabilistic way without exact origins or destination flow restrictions. That is, the model is neither production-constrained nor attraction-constrained, neither doubly-constrained nor unconstrained. Second, choice data has been applied in a novel way to calibrate ATM variables. The paper is organized as follows. In section 2, we briefly discuss the development and characteristics of ATM framework and the model formalization; in section 3 we formalize the spatial choice model and the model estimation method based on observed choice data; section 4 is the case study using practical choice data in retailing. Due to the modifiable areal unit problem (MAUP) in geographical analysis, choice data in two scales (census block group and Zip+4) are discussed and evaluated in detail. We also empirically examine the model results and discuss store scale’s marginal effect that has been observed; and section 5 is the conclusion part and model discussion.

2 A Brief Introduction to Alonso’s Theory of Movement

ATM provides a rather general logical and mathematical framework for spatial interaction models (Alonso 1973, 1978), within which four types of traditional spatial interaction models are represented as special cases (Fotheringham and O'Kelly 1989; Miller and O'Kelly 1991; Wilson 1980) as well as quasi-related properties on marginal effects, such as the relaxed spatial interaction model (Fotheringham and O'Kelly 1989).

The general form of ATM is:

$$P_{ij} = v_i A_i^{\alpha-1} w_j B_j^{\delta-1} d_{ij}^{-\beta}, \tag{1}$$

Where:

$$A_i = \sum_j w_j B_j^{\delta-1} d_{ij}^{-\beta}, \quad i = 1, 2, \cdots, n \tag{2}$$

$$B_j = \sum_i v_i A_i^{\alpha-1} d_{ij}^{-\beta}, \quad j = 1, 2, \cdots, m \tag{3}$$

$$O_i = \sum_j P_{ij} = v_i A_i^{\alpha-1} \sum_j w_j B_j^{\delta-1} d_{ij}^{-\beta} = v_i A_i^{\alpha}, \text{ and} \tag{4}$$

$$Q_j = \sum_i P_{ij} = w_j B_j^{\delta-1} \sum_i v_i A_i^{\alpha-1} d_{ij}^{-\beta} = w_j B_j^{\delta}, \tag{5}$$

In the model, i and j are indices denoting demand and supply zones, respectively; P_{ij} is the unconditional probability of originating in zone i and shopping in alternative j ; O_i is the unconditional probability of being a customer from zone i and Q_j is the unconditional probability of being a customer in destination j ; A_i and B_j are endogenous systemic variables interpreted as the total outside opportunities available to zone i and the total outside competition arrive at store j , respectively (Hua 2001); v_i is the local propulsiveness of zone i and w_j is the site attractiveness of store j , which are correspondingly two sets of observable and measurable attributes inherent to the generation of outflow and attraction of inflow. For example, v_i can be the combination of population size, average household income level and w_j can include stores' floor area or the number of stores in the vicinity; d_{ij} is the spatial separation from zone i to store j , measured by a quantity such as travel distance, travel cost or travel time between zone i and store j ; n and m are the number of origins and destinations, respectively; In this paper n is the number of residential zones and m are the number of stores. β is a distance-decay parameter indicating the level of spatial distance deterrence; α and δ are defined as elasticity variables that denote the rate of outflows or inflows from origin i or to destination j .

Equation (1) is a probabilistic allocation of flows between origin i and destination j depending on their endogenous properties and spatial arrangements. The probabilistic allocation can be used to compare the level of interactions from one origin to multiple destinations, as well as the level of interaction flows from multiple origins to one destination. Equation (2) and (3) are interpreted as relative accessibility of origins and destinations, respectively (Fotheringham and Dignan 1984; Hua 1980). Equation (4) and (5) are sub-models of social and economic movement propulsion and attraction. Large A_i means origin i is relatively accessible to destinations, and large B_j means destination j is relatively accessible to origins. When α or δ changes from zero to one, A_i or B_j will be given more weight in determining the probabilistic allocation, and v_i or w_j will have less influence over the trip allocations. Equation (1) (4) and (5) all need to meet the systemic effects in Equation (2) and (3) (Miller and O'Kelly 1991).

3 Spatial Choice Estimation Using Paired Comparisons

In this research, we attempt to estimate customers' spatial choice probability of store j from a fixed origin i . Given undecided elastic parameters of α and δ , ATM can be applied without the model constraints assumption, which means that the data determine the properties of the model constraints. Paired comparisons method is the technique that induces distinguished choice heterogeneity at both individual and aggregated level consistently (Böckenholt 2001; Stassen, Mittelstaedt *et al.* 1999). Here paired comparisons are employed to calibrate the spatial choice model built from ATM.

Assume there is a matrix **N** with a dimensionality of $I \times J (I = 1,2,\cdots,n; J = 1,2,\cdots,m)$. The spatial choice model is a conditional probability of choosing destinations given fixed origins:

$$P_{j|i} = P_{ij} / O_i = v_i A_i^{\alpha-1} w_j B_j^{\delta-1} d_{ij}^{-\beta} / v_i A_i^{\alpha} = w_j B_j^{\delta-1} d_{ij}^{-\beta} / \sum_k w_k B_k^{\delta-1} d_{ik}^{-\beta}, \tag{6}$$

where $P_{i|j}$ is the conditional probability of originating in zone i, given that the customer patronizes store j. Likewise the conditional probability of store j being chosen by origin i is:

$$P_{i|j} = P_{ij} \big/ Q_i = v_i A_i^{\alpha-1} w_j B_j^{\delta-1} d_{ij}^{-\beta} \big/ w_j B_j^{\delta} = v_i A_i^{\alpha-1} d_{ij}^{-\beta} \Big/ \sum_i v_i A_i^{\alpha-1} d_{ij}^{-\beta}, \tag{7}$$

where $P_{j|i}$ is the conditional probability of patronizing store j, given that the customers live in zone i.

Equation (6) shows that the spatial choice process is only related to the endogenous characteristics of stores (e.g., relative accessibility), the exogenous variables of stores' attractiveness (e.g. floor space), the spatial deterrence between residential zones and stores, and the adjustment factors δ and β that show the degree of the store attraction and spatial separation. However, the propulsiveness and endogenous characteristics of origins show very little about the process of spatial choice. In grocery retailing this means that the choice of a grocery store is only influenced by the characteristics of available stores and the travel cost that the customers can afford. The properties of origins can be reflected by observed expenditure in a store.

Let S_j be exogenous total sales at store j, n_{ij} be an observed expenditure of customers from zone i patronizing store j, and R_j be a weighted adjustment factor used to adjust the observed expenditure. That is, $R_j = S_j \Big/ \sum_i n_{ij}$ measures store sales per unit of sampled sales and $n_{ij} R_j$ then measures the estimated sales in store j from residents of zone i. Compute and take the ratio of a pair of origin-based conditional probabilities as follows:

$$\frac{n_{j|i}}{n_{k|i}} = \frac{n_{ij} R_j \Big/ \sum_m n_{im} R_m}{n_{ik} R_k \Big/ \sum_m n_{im} R_m} = \frac{n_{ij} R_j}{n_{ik} R_k}, \tag{8}$$

Equation (8) is a ratio of a pair of spatial choices observed from the same origin to different destinations. The ratio is determined by the observed expenditures and the adjustment factor of different stores. The observation sampling at stores is itself a stochastic process, and some stores may have more observations than others. More observed customers at a store could simply mean a more determined effort to interview sampled individuals, so it is important to use R_j to eliminate this sampling variability problem and set the benchmark by the stores' sales level.

Theoretically, a model for the ratio of the spatial choice probability derived from Equation (6) is:

$$\frac{P_{j|i}}{P_{k|i}} = \frac{w_j B_j^{\delta-1} d_{ij}^{-\beta} \big/ A_i}{w_k B_k^{\delta-1} d_{ik}^{-\beta} \big/ A_i} = \frac{w_j B_j^{\delta-1} d_{ij}^{-\beta}}{w_k B_k^{\delta-1} d_{ik}^{-\beta}}, \tag{9}$$

Statistically the observational probability equals to the theoretical probability, and we can conclude that $n_{j|i} \big/ n_{k|i} = P_{j|i} \big/ P_{k|i}$. So we have

$$n_{ij} R_j \big/ n_{ik} R_k = w_j B_j^{\delta-1} d_{ij}^{-\beta} \big/ w_k B_k^{\delta-1} d_{ik}^{-\beta}, \tag{10}$$

Linearizing (10), we have

$$\ln \frac{n_{ij} R_j}{n_{ik} R_k} = \ln \frac{w_j}{w_k} + (\delta - 1) \ln \frac{B_j}{B_k} - \beta \ln \frac{d_{ij}}{d_{ik}}, \tag{11}$$

Next formulating equation (5) in terms of B_j^{δ} and linearize:

$$\ln B_j = \frac{1}{\delta}\left(\ln Q_j - \ln w_j\right) = \frac{1}{\delta} \ln \frac{Q_j}{w_j},$$

Therefore, we have

$$\ln\frac{B_j}{B_k}=\ln B_j-\ln B_k=\frac{1}{\delta}\left(\ln\frac{Q_j}{w_j}-\ln\frac{Q_k}{w_k}\right)=\frac{1}{\delta}\ln\left(\frac{Q_j}{Q_k}\cdot\frac{w_k}{w_j}\right), \tag{12}$$

Notice that Q_j can be approximated as market share in grocery retailing. So the ratio of unconditional probability of being a customer in different destinations $\left(Q_j/Q_k\right)$ can be approximated as the ratio of stores' market share, which equals to the ratio of sales levels. So we have $Q_j/Q_k=S_j/S_k$ and Equation (12) becomes:

$$\ln\frac{B_j}{B_k}=\frac{1}{\delta}\ln\left(\frac{S_j}{S_k}\cdot\frac{w_k}{w_j}\right), \tag{13}$$

Substitute Equation (13) to Equation (11):

$$\ln\frac{n_{ij}R_j}{n_{ik}R_k}=\ln\frac{w_j}{w_k}+\frac{(\delta-1)}{\delta}\ln\left(\frac{S_j}{S_k}\cdot\frac{w_k}{w_j}\right)-\beta\ln\frac{d_{ij}}{d_{ik}}, \tag{14}$$

Rewrite (14) for ordinary least square (OLS) regression:

$$\ln\frac{n_{ij}R_j}{n_{ik}R_k}-\ln\frac{w_j}{w_k}=\frac{\delta-1}{\delta}\ln\left(\frac{S_j}{S_k}\cdot\frac{w_k}{w_j}\right)-\beta\ln\frac{d_{ij}}{d_{ik}}, \tag{15}$$

Equation (15) results in an operational linear equation of the form

$$Y=a_1X_1+a_2X_2+\varepsilon, \tag{16}$$

Where, Y, X_1 and X_2 are independent to each other and they are all normally distributed. ε is assumed to be independent of the variables in Equation (16) and follows the normal distribution. So the following identifications can be made:

$$Y=\ln\frac{n_{ij}R_j}{n_{ik}R_k}-\ln\frac{w_j}{w_k},$$

$$X_1=\ln\left(\frac{S_j}{S_k}\cdot\frac{w_k}{w_j}\right),$$

$$X_2=\ln\frac{d_{ij}}{d_{ik}},$$

$a_1=1-1/\delta$, and

$a_2=-\beta.$

Obviously we can solve for the parameters of interest to get:

$\delta=1/(1-a_1)$, and

$\beta=-a_2.$

It has been suggested that ordinary least square (OLS) leads to biased and inconsistent estimators in ATM due to the model's non-linearity and implicitly defined endogenous system variables, and maximum likelihood or instrumental variables should be applied for calibration (Vries, Nijkamp *et al.* 2000). Equation (15) only includes exogenous variables and two parameters, and each of them is independent of the other. In our situation it is reasonable and suitable to employ OLS in spatial choice model estimation. From Equation (5) therefore:

$$B_j = \left(Q_j / w_j\right)^{\frac{1}{\delta}}, \qquad (17)$$

and so, concisely, Equation (6) can be rewritten as:

$$P_{j|i} = w_j^{1/\delta} Q_j^{(\delta-1)/\delta} d_{ij}^{-\beta} \Big/ \sum\nolimits_k w_k^{1/\delta} Q_k^{(\delta-1)/\delta} d_{ik}^{-\beta}, \qquad (18)$$

Where Q_j is the unconditional probability of expenditure patronized to store j, which equals to $S_j \big/ \sum_k S_k$. The insight in (18) is to implicitly define the system variable B_j in terms of the observable variable Q_j for choice prediction and so the prediction model becomes identifiable.

4 Experiments: Two Scales

In this section, choice data are illustrated here for model calibration using Equation (15) (16) and (18) developed in section 3. Data in two scales (Zip+4 and census block group) will be compared here for scale choice and scale sensitivity test through comparison approach.

4.1 Study Area and Observation Data

The study uses data collected about 10 years ago for a grocery chain. The chain has a sufficiently large number of branches in several cities to be suited to providing the kind of controlled trade-off experiment that we need for the purposes of inferring the role of spacing and attraction. For each store in the study region, a sample of customer addresses was derived by geo-coding checks. The amount of data collected from each store varied but because the actual store sales volume at the time was available to the researcher, the representative fraction R_j for each sampled dollar of expenditure could be determined. As is the typical case from the customer spotting literature (Applebaum 1966), each sample dollar is taken to represent $R_j = S/s$ where *S* is the known sum of total weekly store sales and *s* is the sample observed from the customers "spotted." Without revealing the exact data, one can say that this was an exceptionally dense data observation exercise with at least 10% of the sales activity at all the branches of the chain in a mid sized urban metropolitan area accounted for. Clearly more modern data technique associated with frequent shopper cards greatly facilitates the grocery operations analysis of the trade area of particular stores. Our goal is to extract some generalized spatial choice parameters from the choice based information gathered by "customer spotting."

4.2 Scale Problem and Data Pre-processing

Scale and unit specification in geographical analysis have been widely addressed and identified as modifiable areal unit problem (MAUP) (Cao and Lam 1997; Openshaw and Taylor 1981). The choice of different spatial scales and representation units could detect randomly various spatial choice patterns in retailing (Bailey and Gatrell 1995). MAUP is rooted in scale-dependency of geographical phenomena and different ways in eliminating contextual differences, place and individual heterogeneity in data aggregation (Jones and Duncan 1996). The manipulation of spatial data representation causes the consequence of questionable validity of analytical results (Murray and Weintraub 2002). Researchers proposed new approaches, such as frame independent spatial analysis approaches to eliminate MAUP effects to give consistent results (Tobler 1989); or choose appropriate models that are less susceptible to MAUP effects (Murray and Weintraub 2002). The choice of an appropriate geographical scale is usually determined by research objective, the phenomenon under study, and the data that has been collected. Tobler (1989) suggested that if the analysis results are sensitive to MAUP effects, typical way is to use the most disaggregated information. Basically, the choice of analysis scale and unit is an empirical process (Bailey and Gatrell 1995).

In this research 39,675 check records have been collected and may be aggregated to two scales: census block group (Q1) and Zip+4 (Q2). These choice data belong to 878 block groups and 11,345 Zip+4 areas. Observed choice data will be aggregated by their origins and destinations: all the observed choices from the same origin to the same store will be aggregated as one observed spatial flow. Typically, for aggregation in Zip+4 scale, the spatial extent for each Zip+4 region is so small that we can treat observed travel distances from the same Zip+4 origin to the same store as approximately equal. For block group scale aggregation, the travel distances may vary significantly from within the origin to the same destination, and so the aggregated travel distance is adjusted by expenditure observed:

$$D_{ij} = \sum_k y_{ijk} d_{ijk} \Big/ \sum_k y_{ijk} \,,$$

Here D_{ij} is weighted distance from block group i to store j ; y_{ijk} is expenditure of customer k within block i at store j; and d_{ijk} is travel distance that the customer k within block i needs to cover to store j .

Another problem in the choice data aggregation is that there exist many combinations of spatial choices with zero expenditure even after aggregation. These observed gaps in spatial choices at certain stores indicate infeasible choices or missing data during the choice data collection. This problem becomes particularly sharp for Zip+4 because at Zip+4 it easily happens that there is only one choice observed from the origin. For each census block group, more choices are generally observed at this more aggregated scale, which makes choice observations with zero expenditure less likely to happen. The aggregated choices with zero expenditure are discarded in analysis.

The observed choices among individuals over discrete alternative stores are compared in a combinational way to pool the data of spatial choice differences from the same origin to different destinations. The comparisons of choices reflect spatial preferences caused by blended exogenous and endogenous determinants. Choice comparison pairs are generated for model estimation in two aggregation scales separately. To generate comparison pairs, one requirement is that there are at least two aggregated choices observed to generate one comparison pair. If there is only one aggregated choice observed for a Zip+4 region or census block group, there is no way to acquire paired comparisons for this single-choice origin and this aggregated observation will be discarded as ineffective. We are interested in the information contained in those situations where there is some kind of split in the chosen alternatives: often these will be pronounced in areas situated within easy reach of multiple branches. The more aggregated choice observations from one region, the richer the paired spatial choice comparisons. The number of paired comparisons from one origin to *n* destinations is *n* choose 2 (C_n^2). In our choice data, block group aggregation produces 11,897 paired comparisons while Zip+4 scale only produces 8,047 paired comparisons.

In *Fig. 1* the horizon axis is categorized by the aggregated choice frequencies, and the vertical axis is its corresponding percentage for each frequency group. For Zip+4 there are 11,345 origins, among which 8,078 origins have only one aggregated choice reaching a high percentage of 71.2% (8,078/11,345); for census block group, there are 878 block groups in total, and the number of block groups with only one aggregated choice is 201, which is 22.9%

(201/878) of the total block groups.

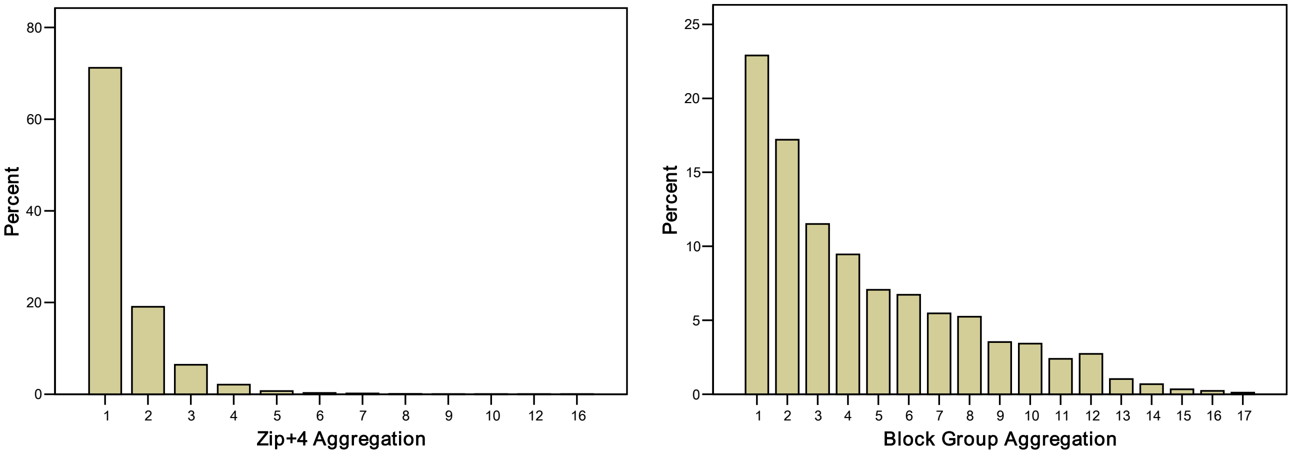

Fig. 1 Choice Aggregation Comparison in Two Scales

Though generally the most disaggregated information should be use to reduce MAUP effects in spatial analysis (Murray and Weintraub 2002; Tobler 1989), more aggregated scale (census block group) has included more paired comparison information and consequently generates better results in this problem.

4.3 Model Output

The paired choice comparisons in two spatial aggregations have been generated for OLS regression. *Table 2* and *Table 3* show the regression results. In both scales, the dependent and independent variables approximately follow normal distribution and the error also follows the normal distribution. From equations in (16) the parameters are:

$\delta = 1/(1-a_1)$, and

$\beta = -a_2$.

So we can calculate δ and β based on regression coefficients. In both scales, all the model coefficients are statistically significant. δ in both scales are found to be near 0.6. Since α is unknown, the properties of system production is unknown and the probability of spatial choice does not relate to production properties. So ATM in this problem is at least a quasi-attraction-constrained spatial interaction model. β in Zip+4 aggregation is slightly greater than 1 meaning that the distance-deterrence changes linearly with travel cost (travel distance or time); while in block group aggregation, β is approximately equal to 2 meaning the distance-deterrence increases with travel cost square. Travel cost has dominant influence over customers' spatial choices in block group level. Adjust R-square in block group scale (0.587) is much higher than in Zip+4 scale (0.345), which coincides with our preliminary scale choice analysis that the more aggregated data (census block group) produces better fitted model estimators than more disaggregated data (Zip+4).

Table 2: Regression Using Data in Zip+4 Aggregation

Source	SS	df	MS	
Model	9519.63	2	4759.813	F(2, 8045)=2123.757
Residual	18030.64	8045	2.241	Sig. F Change=.000
Total	27550.27	8047		Adjusted R-Square=.345

Variable	B	Std. Err	t	Sig.	Mean
LnX1	-.600	.046	-13.129	.000	.0009
LnX2	-1.066	.017	-64.435	.000	-.2212
YY					-.0034
Parameters		$\delta = 0.625$		$\beta = 1.066$	

Table 3: Regression Using Data in Census Block Group Aggregation

Source	SS	df	MS	
Model	43336.56	2	21668.281	F(2, 11895)=8448.510
Residual	30507.65	11895	2.565	Sig. F Change=.000
Total	73844.22	11897		Adjusted R-Square=.587

Variable	B	Std. Err	t	Sig.	Mean
LnX1	-0.673	.041	-16.308	.000	.0146
LnX2	-2.160	.017	-129.986	.000	-.3321
YY					.5916
Parameters		$\delta = 0.598$		$\beta = 2.160$	

According to Equation (18), when at census block group scale the prediction of conditional spatial choice is:

$$\begin{aligned} P_{j|i} &= w_j^{1/\delta} Q_j^{(\delta-1)/\delta} d_{ij}^{-\beta} \Big/ \sum\nolimits_k w_k^{1/\delta} Q_k^{(\delta-1)/\delta}\, d_{ik}^{-\beta} \\ &= w_j^{1/0.598} Q_j^{(0.598-1)/0.598} d_{ij}^{-2.160} \Big/ \sum\nolimits_k w_k^{1/0.598} Q_k^{(0.598-1)/0.598}\, d_{ik}^{-2.160} \\ &= w_j^{1.673} Q_j^{-0.673} d_{ij}^{-2.160} \Big/ \sum\nolimits_k w_k^{1.673} Q_k^{-0.673}\, d_{ik}^{-2.160}, \end{aligned} \qquad (19)$$

In Equation (19), w_j and Q_j are two different measures denoting store attractiveness using store floor space and store sales, respectively. Since a floor space increase prompts store sales to increase and both are non-linearly interrelated, we can infer that the negative effect of sales is offset somewhat by the positive contribution of floor space. When the floor space increases, the spatial choice probability decreases. The bigger the floor space, the greater the cancellation of the floor space's contribution to spatial choice probability. This contradicts the assumption that bigger floor space denotes higher store attractiveness and tends to attract more patronage. In Equation (19), an exploratory data analysis helps to reveal the mechanism of variables' nonlinear change in *Fig. 2*.

Fig. 2 shows two graphs about the relationship between store floor space and store sales, and between store floor space and system variable B_j. A total of 118 real stores' sales and their corresponding B_j values are plotted in ascending order of store floor space. From the graph we can see that floor space increase introduces the store sales increase. Suppose the following exponential nonlinear relationship exists between store floor space w_j and store sales level Q_j (Mahajan, Sharma *et al.* 1988; O'Kelly 2001):

$$Q_j = f(w_j), \tag{20}$$

The rate of floor space promotion to store sales level can be represented as the first derivative $f' = \partial Q / \partial w$. Equation (20) may be represented using an exponential tendency line in the first graph of *Fig*. 2. The marginal effect of store floor space is observed that f' gradually decreases to zero with the increase of store floor space, meaning that the increase rate of store sales level promoted by per sq. foot of floor space increase gradually diminishes. The floor area promotes stores' attraction when the stores' size is around a certain scale range. When the floor space exceeds a certain scale, the promotional effect of store floor space diminishes rapidly. This marginal effect of floor space is reflected by store sales in this model. Store sales level Q_j acts as a balancing factor with a negative exponent to adjust the choice probability overestimated by store floor space w_j in equation (19). This implies that the increase of floor space can never be an unbounded way to improve store's performance and the sales per sq. foot eventually decreases. From the perspective of the trade area, each store in space has its own spatial influence in its trade area and the trade area expands when the store floor space increases. Each unit of store floor space increase introduces trade area expansion. However, the trade area expansion rate decreases due to the marginal effect of cumulative floor space increase.

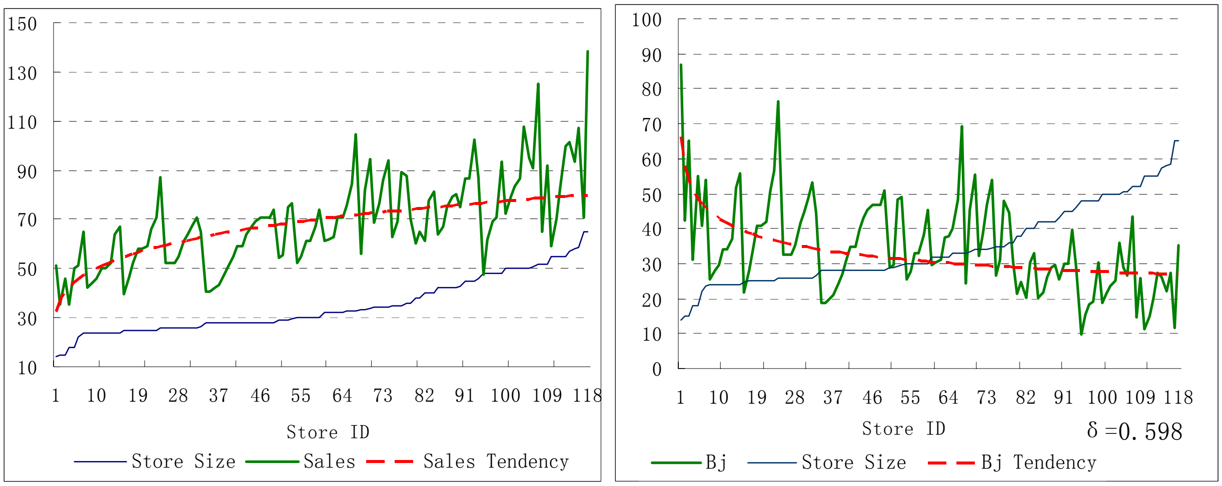

Fig. 2 Tendency Exploration of Store Sales and B_j *with the increasing Store Size Change*

B_j reflects a blended measure of exogenous and endogenous variables, such as site relative accessibility, sales per sq. foot, and so on. Equation (17) represents B_j as a function of the ratio between store sales Q_j and store floor space w_j, as well as the elasticity variable δ. B_j is an indicator of stores' performance when δ is fixed, which is easier to be measured and interpreted compared with the idea of relative accessibility. The exponential tendency line of B_j is shown in the second graph in *Fig*. 2. As has been discussed above, the design and provision of appropriate grocery store size is vital for the stores' survivability under market competition, and moderate store size helps to retain its competition power. Store floor space versus sales share also has been discussed in (O'Kelly 2001).

5 Conclusion

ATM provides a general framework of movement in a spatial context and establishes an important advance in Spatial Interaction Modeling (O'Kelly 2004; Vries, Nijkamp *et al.* 2000). In the past decade, researchers have devoted themselves to model calibration and interpretation using different model calibration techniques. However, model calibration difficulties as mentioned in introduction part still exist and thus deter the wide application of ATM. This paper is positioned with two significant contributions: it takes ATM as a framework for building spatial choice model and applies choice-based samples for ATM parameter calibration. Choice data are used in a novel way to calibrate ATM and the data results in a quasi-attraction-constrained SI model. Different from previous practices, ATM avoids flow constraint assumptions (production-constrained or doubly-constrained, etc.) in spatial choice modeling. Components related to spatial choice probability also have been investigated: stores' endogenous characteristics (store performance), exogenous factors (floor space) and the spatial deterrence (travel costs). More aggregated data produces better results than more disaggregated data due to more paired comparisons at more aggregated scale. We also find out that store's performance is tightly related with the store's scale (attractiveness in general).

There are also some limitations in this research. Basically, through the construction of spatial choice model, we successfully calibrated δ and β, but α is still unknown and needs further detection. In addition, the choice model built here still only considers distance and facility properties due to the data limitation, which ignores socioeconomic characteristics of travelers and some other characteristics of grocery stores. Also grocery shopping trip is taken as single-purpose trip, while in reality most shopping trips are multi-purpose (Hanson 1980; Leszczyc, Sinha *et al.* 2004; O'Kelly 1981) and shoppers are likely to choose clustered or agglomerated stores instead of nearer ones. We simply use store's floor space to denote the store's attractiveness which is also determined by many other variables, such as the number of checkout counters, the number of nongrocery products sold, double or triple manufacturer's coupons, if the store is open 24 hours, number of similar stores located in the vicinity, if the bank is available, etc. It is important to call attention to modeling of the facility's "attractiveness" and segmented shoppers in real application.

Acknowledges

The authors are grateful to Dr. Changshan Wu and Dr. Ningchuan Xiao for their comments and revision of the earlier manuscript.

References

Alonso W., 1973. *Notional Interregional Demographic Accounts: A Prototype*, Monograph 17, Institute of Urban and Regional Development, University of California, Berkeley.

Alonso W., 1978. "A Theory of Movement". *in Human Settlement Systems*. N. M. Hansen. Cambridge, Ballinger. 197-211.

Anselin L., 1982. "Implicit Functional Relationship Between Systemic Effects in a General Model of Movement." *Regional Science and Urban Economics* **12** 365-380.

Applebaum W., 1966. "Methods for Determining Store Trade Areas, Market Penetration and Potential Sales." *Journal of marketing Research* **3** 127-141.

Bailey T. C. and A. C. Gatrell, 1995. *Interactive Spatial Data Analysis*.New York, Longman Scientific & Technical.

Böckenholt U., 2001. "Hierarchical Modeling of Paired Comparison Data." *Psychological Methods* **6** 49-66.

Cao C. and N. S.-N. Lam, 1997. "Understanding the Scale and Resolution Effects in Remote Sensing and GIS". *in Scale in Remote Sensing and GIS*. D. A. Quattrochi and M. F. Goodchild. New York, Lewis Publishers. 57-72.

Cosslett S. R., 1981. "Maximum Likelihood Estimator for Choice-Based Samples." *Econometrica* **49** 1289-1316.

De VOS A. F. and J. A. Bikker, 1982. "Interdependent Multiplicative Models for Allocation and Aggregates: A Generalization of Gravity Models". Rep. No. IAWE-80, Vrije Universiteit, Amsterdam.

Fischer M. M., 2000. "Spatial Interaction Models and the Role of Geographic Information Systems". *in Spatial Models and GIS*. A. S. Fotheringham and M. Wegener. London, Taylor & Francis. 33-44.

Fotheringham A. S. and T. Dignan, 1984. "Further contributions to a general theory of movement." *Annals of the Association of American Geographers* **74** 620-633.

Fotheringham A. S. and M. E. O'Kelly, 1989. *Spatial Interaction Models: Formulations and Applications*.Boston, Kluwer Academic Publishers.

Getis A., 1991. "Spatial interaction and spatial autocorrelation: A cross-product approach." *Environment and Planning A* **23** 1269-1277.

Hanson S., 1980. "Spatial diversification and multipurpose travel: Implications for choice theory." *geographical Analysis* **16** 244-249.

Hua C.-I., 2001. "Alonso's Systemic Model: A Review and Representation." *International Regional Science Review* **24** 360-385.

Hua C., 1980. "An Exploration of the Nature and Rationale of a Systemic Model." *Environment and Planning A* **12** 713-726.

Imbens G. W., 1992. "An Efficient Method of Moments Estimator for Discrete Choice Models With Choice-Based Sampling." *Econometrica* **60** 1187-1214.

Jones K. and C. Duncan, 1996. "People and Places: the Multilevel Model as a General Framework for the Quantitative Analysis of Geographical Data". *in Spatial Analysis: Modelling in a GIS Environment*. P. Longley and M. Batty. New York, John Wiley and Sons. 79-104.

Lancaster T., 1997. "Bayes WESML Posterior inference from choice-based samples." *Journal of Econometrics* **79** 291-303.

Ledent J., 1980. "Calibrating Alonso's General Theory of Movement: The Case of Interprovincial

Migration Flows in Canada." *Sistemi Urbani* **2** 327-358.

Leszczyc P., A. Sinha and A. Sahgal, 2004. "The effect of multi-purpose shopping on pricing and location strategy for grocery stores." *Journal of Retailing* **80** 85-99.

Mahajan V., S. Sharma and R. Kerin, 1988. "Accessing Market Penetration Opportunities and Saturation Potential for Multi-Store, Multi-Market Retailers." *Journal of Retailing* **64** 315-333.

Manski C. and S. Lerman, 1977. "The Estimation of Choice Probabilities from Choice-Based Samples." *Econometrica* **45**.

Manski C. and D. McFadden, 1981. "Alternative Estimators and Sample Designs for Discrete Choice Analysis". *in Structural Analysis of Discrete Data*. C. F. Manski and D. McFadden. Cambridge, Massachusetts, M.I.T. Press.

Miller H. J. and M. E. O'Kelly, 1990. "Incorporating situational effects into retail market area delimitation." *The Operational Geographer* **8** 9-12.

Miller H. J. and M. E. O'Kelly, 1991. "Properties and estimation of a production-constrained Alonso model." *Environment and Planning A* **23** 127-138.

Mueser P. R., 1989. "Measuring the Impact of Locational Characteristics on Migration: Interpreting Cross-Sectional Analyses." *Demography* **26** 499-513.

Murray A. T. and A. Weintraub, 2002. "Scale and Unit Specification Influences in Harvest Scheduling with Maximum Area Restrictions." *Forest Science* **48** 779-789.

O'Kelly M., 1981. "A model for the demand of retail facilities, incorporating multistop, multipurpose trips." *Geographical Analysis* **13** 134-148.

O'Kelly M., 1999. "Trade-area models and choice-based samples: methods." *Environment and Planning A* **31** 613-627.

O'Kelly M., 2004. "Isard's contributions to spatial interaction modeling." *Journal of Geographical Systems* **6** 43-54.

O'Kelly M., 2005. "Parameter Estimation in Choice Based Models for Retail Trade Area using OLS." *in preparation*.

O'Kelly M. E., 2001. "Retail market share and saturation." *Journal of Retailing and Consumer Services* **8** 37-45.

Openshaw S. and P. J. Taylor, 1981. "The Modifiable Areal Unit Problem". *in Quantitative Geography: A British View*. N. Wrigley and R. Bennett. London, Routledge and Kegan Paul. 60-69.

Poot J., 1986. "A System Approach to Modeling the Inter-urban Exchange of Workers in New Zealand." *Scottish Journal of Political Economy* **33** 249-274.

Porell F. W. and C. Hua, 1984. "An Econometric Procedure for Estimation of a Generalized Systematic Gravity Model under Incomplete Information about the System." *Regional Science and Urban Economics* **11** 585-606.

Roy J. R. and J.-C. Thill, 2004. "Spatial Interaction Modelling." *Regional Science* **83** 339-361.

Stassen R. E., J. D. Mittelstaedt and R. A. Mittelstaedt, 1999. "Assortment overlap: its effect on shopping patterns in a retail market when the distributions of prices and goods are known." *Journal of Retailing* **75** 371-386.

Tabuchi T., 1984. "The Systemic Variables and Elasticities in Alonso's General Theory of Movement." *Regional Science and Urban Economics* **14** 249-264.

Thill J.-C. and J. L. Horowitz, 1991. "Estimating a Destination-Choice Model from a Choice-Based Sample with Limited Information." *Geographical Analysis* **23** 298-315.

Tobler W. R., 1989. "Frame Independent Spatial Analysis". *in The Accuracy of Spatial Database*. M. F. Goodchild and S. Gopal. New York, Taylor and Francis. 115-122.

Vries J. J. D., P. Nijkamp and P. Rietveld, 2000, "Alonso's General Theory of Movement." Tinbergen Institute Discussion Paper http://www.tinbergen.nl/discussionpapers/00062.pdf *Accessed on 12/13/2004*.

Vries J. J. D., P. Nijkamp and P. Rietveld, 2002. "Estimation of Alonso's Theory of Movements by Means of Instrumental Variables." *Network and Spatial Economics* **2** 107-126.

Waldman D. M., 2000. "Estimation in Discrete Choice Models With Choice-Based Samples." *The American Statistician* **54** 303-306.

Williams P. A. and A. S. Fotheringham, 1984. "The Calibration of Spatial Interaction Models by Maximum Likelihood Estimation with Program SIMODEL". *in Geographic Monograph Series*, Department of Geography, Indiana University. **7**.

Wilson A. G., 1980. "Comments on Alonso's 'Theory of Movements'." *Environment and Planning A* **12** 727-732.