ROLAND T. RUST and NAVEEN DONTHU*

No retail store choice model, no matter how many relevant variables it might include, can realistically expect to model *all* the variation in store choice. There are always some variables that are left out, because they are difficult to measure, they have not yet been conceptualized in theory, or their estimated parameter stability suffers when an excessive number of predictors are included. Because these omitted variables can be correlated with geographic location, model misspecification error may itself be correlated with location. Estimating the geographically localized misspecification errors therefore suggests itself as a method for estimating (and predicting) the effects of these omitted variables. The authors show that spatial nonstationarity of the model parameters may also be expressed as an instance of omitted variables and therefore be addressed using their method. They show, using both a simulation study and an empirical natural experiment, that estimating the geographically localized misspecification error can appreciably reduce prediction error, even when the predictor model is reasonably well specified.

Capturing Geographically Localized Misspecification Error in Retail Store Choice Models

Selecting the best possible site for a retail store (or service facility) is often a critical factor in that store's eventual success (Achabal, Gorr, and Mahajan 1982; Ghosh and Craig 1983). The site selection problem is more complicated for a chain, which must worry about whether the new store will cannibalize its existing stores by stealing their customers. In either case, site selection cannot be accomplished effectively without knowledge of where the customers are and how they choose their store, as well as knowledge of the geographic configuration of existing stores.

Because the geographic configuration of the existing stores is easily obtained, the most important information that must be determined by primary research is where the customers are located geographically and how they choose a store.

Methods for flexibly estimating the geographic customer density from sample scatter plot data have been proposed in recent years (Donthu 1991; Donthu and Rust 1989; Rust and Brown 1986). There is also extensive literature on consumer choice models, much of it based on application of logit (e.g., Batsell and Lodish 1981; Gensch and Recker 1979; Guadagni and Little 1983; McFadden 1980). Consumer choice models, when applied to retail attraction, have been in common usage for many years (e.g., Fotheringham 1980; Huff 1964; Louviere 1984).

Retail attraction models typically predict store choice based on a number of variables. However, it is a practical impossibility to include *all* of the variables that affect choice. There are several reasons for this. First, some variables may be very difficult to measure, and thus they are not practical to include. Second, some variables that affect choice may not have been conceptualized or identified by the researcher. Third, even if it were possible to identify and measure all relevant predictors, it would not be advisable, because the use of too many variables leads to parameter instability and a decline in predictive accuracy (Inagaki 1977; Rust and Schmittlein 1985).

The result is that a well-specified retail attraction model will include what are thought to be the most important variables, but it will necessarily omit many others. The inevitable presence of omitted variables means that the typical model is misspecified (in this case, underspecified). This result is *inevitable*, and the most careful model building and testing will never be able to completely specify the model without suffering parameter instability and declines in predictive accuracy, which will more than offset the benefits.

*Roland T. Rust is Professor of Marketing and Director of the Center for Services Marketing, Owen Graduate School of Management, Vanderbilt University. Naveen Donthu is Associate Professor of Marketing, Georgia State University.

Journal of Marketing Research
Vol. XXXII (February 1995), 103–110

In Inagaki's terms, the total prediction error would increase, because the increase in estimation error would more than offset the decrease in modeling error. Thus, the intelligent researcher resigns himself or herself to some degree of model misspecification and the existence of omitted variables.

The effects of these omitted variables can be correlated with geographic location. The idea that aspects of retail attraction models may vary geographically is not new. Many researchers have shown that the model parameters may suffer from spatial nonstationarity (e.g., Bucklin 1971; Craig, Ghosh, and McLafferty 1984; Fotheringham 1981; Ghosh 1984).

In our case, we are observing not only that the *model parameters* may vary by geographical location but also that the *effects of the omitted variables* may vary by geographical location. We will show that estimating the localized effects of omitted variables simultaneously addresses the issue of parameter nonstationarity.

By estimating the localized effects of the omitted variables, which is essentially equivalent to estimating the localized misspecification errors, we hope to reduce the model prediction error. Subsequent sections provide evidence that this approach works, but first we give a simplified example that shows conceptually how capturing misspecification error geographically can produce benefits.

For example, suppose there are two grocery chains, A and B, and that chain A sells primarily Mexican groceries, whereas chain B sells primarily Chinese groceries. For purposes of simplicity, let us suppose further that our retail attraction model includes only distance. Consider two neighborhoods that are equidistant from a store belonging to chain A and a store belonging to chain B.

According to the attraction model, equal probabilities (.5) of store choice would be predicted for both neighborhoods. However, suppose one neighborhood is primarily Mexican, whereas the other is primarily Chinese. The Mexican neighborhood, because of the unobserved (not specified) predictor variable—ethnic similarity—would probably tend to choose chain A; the other neighborhood would tend to choose chain B.

In this case, it is clear that the impact of the misspecification of the choice model is distributed geographically. If we can capture this geographic distribution in model misspecification error, we may be able to improve the grossly misspecified choice model *even without adding predictor variables*. Of course, we will first specify the model as well as possible. However, even a good model leaves out variables that may affect choice.

We will demonstrate that it is possible to capture these geographically localized misspecification errors, taking advantage of the effects of the omitted variables that can be correlated with geographic location.

In summary, we hypothesize that the misspecification of retail store choice models can be geographically localized. We test this hypothesis and show how this geographic component can be captured using nonparametric density estimation. Section 2 provides a method for estimating geographically localized model misspecification error, whereas section 3 provides the results of validating tests of the model based on simulated data. An empirical validation test based on before-after store choice is reported in section 4, and conclusions and implications are summarized in section 5.

THE MODEL

We first discuss the relationship between model misspecification, utility, and choice. We then show how the proposed model (to capture geographically localized model misspecification errors) can be operationalized by combining the methodologies of logit regression and nonparametric density estimation.

Utility and Choice

To facilitate exposition, we assume a very simple model for individual store choice. Many other choice frameworks might also be assumed, and they would result in changes in the specifics (but not the general direction) of the development. Specifically, we assume a Luce choice rule (Luce 1959). This simple rule is victim to the IIA property, which has stimulated researchers to propose more complicated alternative models (Batsell and Polking 1985; Currim 1982; Fotheringham 1985; Kamakura and Srivastava 1984). Nevertheless, this simple model is entirely adequate to demonstrate the concept of geographically localized misspecification error.

We assume that the utility U_{ij} of chain j to individual i is

$$U_{ij} = \exp[\alpha\, d_{ij} + Y_{ij}\,\beta + \delta_{ij}] \tag{1}$$

where

d_{ij} = distance from individual i to the nearest store in chain j

α = a coefficient on distance, assumed to be homogeneous across individuals

Y_{ij} = a vector of predictor variables other than distance

β = a vector of coefficients, assumed to be homogeneous across individuals

δ_{ij} = a random error term, assumed to be distributed extreme value

and that the probability of individual i choosing chain j is

$$P_{ij} = U_{ij} / \sum_{k} U_{jk}. \tag{2}$$

Subsequently, we will estimate the parameters of this model using logit regression.

We now recognize that the error term, δ_{ij}, can be broken down into two components: deviation based on the impact of omitted predictor variables and deviation due to random sampling error. Note that the impact of omitted predictors is what we are calling misspecification error.[1] In other words,

$$\delta_{ij} = ME_j(X_i) + \varepsilon_{ij} \tag{3}$$

with

$ME_j(X_i)$ = model misspecification error at point X_i

[1]A linear model with enough variables may fit even a nonlinear functional relationship to any requested level of precision if we include polynomial terms. Thus, any model misspecification can be viewed as the omission of necessary predictor variables.

ε_{ij} = random sampling error, assumed to have mean 0 but with no specific distribution assumed.

We are assuming that, observed across data points, the combined distribution of model misspecification error (residuals of the unobserved variables) plus sampling error is extreme value. This is not the same as assuming ε_{ij} is distributed extreme value at any data point. In fact, we require no strong distributional assumptions on ε_{ij}.

Omitted Variables and Parameter Nonstationarity

Although it may not be immediately obvious, this approach also addresses the issue of parameter nonstationarity (parameters differ across geographic areas). To illustrate, consider a simple example. Let us suppose that there is only one predictor variable (besides distance) in the model, and that its parameter differs across two geographic areas, 1 and 2. Let us suppose that the utility expression is of the form:

$$U_{1ij} = \exp[\alpha\, d_{ij} + Y_{ij}\, \beta_1 + \delta_{ij}] \quad (4)$$

$$U_{2ij} = \exp[\alpha\, d_{ij} + Y_{ij}\, \beta_2 + \delta_{ij}]$$

where U_{1ij} and U_{2ij} are the utilities in regions 1 and 2 respectively, and β_1 and β_2 are the nonstationary parameter values.

The parameter nonstationarity inherent in the above equations can be expressed as an instance of omitted variables and thereby reduced to the situation we have already addressed. We will do this by defining (omitted) variables, which involve dummy variables for the different geographic regions. We define a dummy variable, D, to be equal to 1 if the individual resides in region 1 and equal to 0 otherwise. We then define a variable Z_{ij}, which is the product of D_i and Y_{ij}. The equation, fully specified, would then include d_{ij}, Y_{ij}, and Z_{ij}.

$$U_{ij} = \exp\,[\alpha\, d_{ij} + Y_{ij}\, \beta + \gamma\, Z_{ij} + \delta_{ij}] \quad (5)$$

If there were no measurement error, the estimated coefficients would then equal β_2 for $\hat{\beta}$, and $(\beta_1 - \beta_2)$ for $\hat{\gamma}$. With measurement error, the estimated coefficients would have the usual properties (e.g., unbiasedness and consistency) ascribed to the estimated coefficients of fully-specified regression models.

Although we show only a simple example, the extension to multiple regions is straightforward. Of course, it would be very difficult to know in advance which regions would result in different parameter values. That, of course, is yet another reason why capturing this parameter nonstationarity implicitly by estimating the geographically localized misspecification errors may be appealing.

Estimation

We use a multistage approach to estimating the model parameters. (The complexity of the model makes simultaneous estimation a practical impossibility.) First, we estimate the model in equation (1) using conventional maximum likelihood multinomial logit estimation. This yields estimates for α and β. Second, we use the estimated logit parameters in conjunction with nonparametric estimates of the customer densities for each chain to infer the geographically localized model misspecification error. The steps required in the second stage are explained subsequently.

For simplicity of exposition, let us consider a market in which there are two competing chains: chain j and chain k. (Generalization to more than two chains is straightforward.) Let S_j and S_k ($= 1 - S_j$) be the proportional shares of chains j and k, respectively.

We use nonparametric density estimation to estimate the customer density for each chain. Following Donthu and Rust's (1989) study, we estimate chain j's customer density at point X as:

$$\hat{f}_j(X) = (n_j h^2)^{-1} \sum K[(X - X_i) / h] \quad (6)$$

where X refers to geographic location, X_i is the location of individual i, n_j is the number of customers in the sample who choose chain j, h is the smoothing factor, and K is the kernel function. The kernel function K is chosen to be a symmetric distribution, such as a rectangular distribution or a bivariate normal distribution. The remainder of our exposition is independent of the choice of h and K. For our examples we use a bivariate normal kernel function (with independent margins of equal variance) and an automatically chosen h value (Silverman 1986, p. 86.).

We are assuming that distance is determined with respect to a Euclidean metric. The estimated density is unaffected by the choice of the origin or unit of distance. However, the North-South and East-West axes must use the same unit of distance.

We would like to use the sample data to estimate the probability of choosing chain j (or k) at any point in the space. The densities obtained from equation (6) give us the relative customer densities for each chain at any point. If we weight those densities by sample size S_j, we can obtain the estimated probability of choosing a chain at any point in the space. For example, the probability of a respondent at point X_i choosing chain j is estimated to be:

$$P_j(X_i) = S_j\hat{f}_j(X_i) / (S_j\hat{f}_j(X_i) + S_k\hat{f}_k(X_i)) \quad (7)$$

$$= [1 + (S_k/S_j)\,(\hat{f}_k(X_i) / \hat{f}_j(X_i)]^{-1}$$

But equation (7) can also be linked to the logit parameters. From equations (1), (2), and (3), we obtain:

$$P_j(X_i) = \{1 + \exp[\alpha(d_k(X_i) - d_j(X_i)) + (Y_{ik} - Y_{ij})\beta \quad (8)$$

$$+ (ME_k(X_i) - ME_j(X_i)) + (\varepsilon_k - \varepsilon_j)]\}^{-1}$$

where Y_{ik} and Y_{ij} are the predictor variable values for individual i for chain k and j respectively.[2] Therefore, after some algebra and combining with equation (7),

$$\Delta_{jk}(X_i) = (ME_j(X_i) - ME_k(X_i)) \quad (9)$$

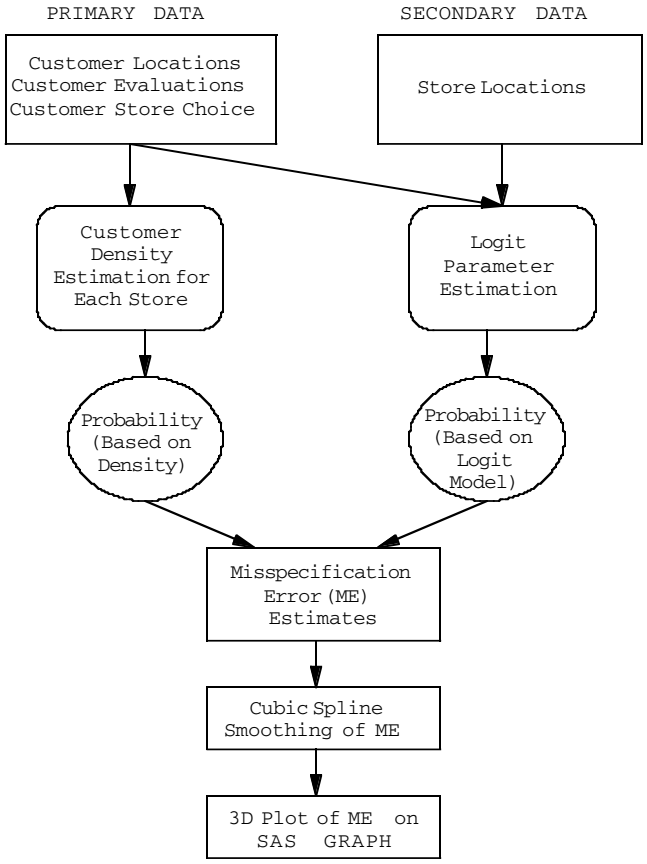
$$= \ln(S_j\hat{f}_j(X_i) / S_k\hat{f}_k(X_i)) + \alpha(d_j(X_i) - d_k(X_i))$$

$$+ (Y_{ij} - Y_{ik})\beta + (\varepsilon_j - \varepsilon_k)$$

Noting that $E(\varepsilon_j - \varepsilon_k) = 0$ and that all other terms on the right-hand side have been estimated, we can use those estimates (ignoring the $e_j - e_k$ term, which is approximately

[2]If a chain was not rated by an individual, we substituted the mean rating across those who rated the chain.

Figure 1
ESTIMATION AND PLOTTING OF MISSPECIFICATION ERROR



zero) to approximate $\Delta_{jk}(X_i)$, the difference between the two chains' misspecification errors:

$$(10)\quad \hat{\Delta}_{jk}(X_i) = \ln(S_j\hat{f}_j(X_i) / S_k\hat{f}_k(X_i)) + \hat{\alpha}(d_j(X_i) - d_k(X_i)) + (Y_{ij} - Y_{ik})\hat{\beta}$$

This gives us estimates of misspecification error differences for the points in space that correspond to individual locations in the sample, but we also want to obtain estimates for all geographic locations. We do this by spline smoothing using the G3GRID procedure in SAS GRAPH (SAS Institute 1987). The major steps involved in estimating the misspecification error is shown in flow chart form in Figure 1.

We cannot approximate any individual chain's geographically localized misspecification error but only the differences between chains. The units are arbitrary and reflect that the logit regression units are also of an arbitrary scale (Swait and Louviere 1993).

Properties of the Estimators

The estimators $\hat{\alpha}$ and $\hat{\beta}$ assume all the favorable properties of logit coefficients that are estimated by maximum likelihood (asymptotic efficiency, consistency, etc.). Likewise $\hat{f}$ assumes the favorable properties of kernel nonparametric density estimators (consistency, uniform consistency). The number S_j, which chooses a particular chain is binomially distributed. It is unbiased, efficient, and consistent, as is any binomial estimator.

The accuracy of the approximation of $\Delta_{jk}(\underline{X})$ is harder to assess, because it results from an expression involving other estimators whose distributions are not tractably combinable, and anticipated values of the predictors Y. Viewed as an estimator, we know that $\hat{\Delta}$ is consistent (if Y_k and Y_j are known), because as the sample size goes to infinity, the error distribution of each term collapses on zero. However, we know little about its efficiency. Thus, we must assess the usefulness of the approximation through simulations and empirical tests. The methodology for these simulations and empirical tests is outlined in the next two sections.

VALIDATION BY SIMULATION

We first must establish that the proposed approach can estimate geographically localized misspecification if it exists. We also must explore the model's ability to predict market shares after a retail chain expands to a new location. We examine the predictive accuracy of the model compared to that of two simpler models, one that models misspecification and one that does not.

Competing Models

Uniform misspecification model. The first competing model is a uniform misspecification model (UMM), which assumes that model misspecification exists but that the effect of these unobserved variables is constant through geographic space. Mathematically, the UMM model replaces equation (1)'s expression for the utility of chain j to individual i by:

$$(11)\quad U_{ij} = \exp[\alpha d_{ij} + Y_{ij}\beta + \psi_j + \delta_{ij}],$$

where α, d_{ij}, Y_{ij}, and β are defined as before, ψ_j is a chain-specified intercept term that reflects chain j's chain-specific effect, and δ_{ij} is still a distributed extreme value but is now entirely comprised of random error.

Simple logit model. The second competing model is a simple logit model (SL), which assumes that model misspecification does not exist. Hence, it is similar to equation (1), with δ_{ij} wholly comprised of random error. It can be mathematically expressed as:

$$(12)\quad U_{ij} = \exp[\alpha d_{ij} + Y_{ij}\beta + \delta_{ij}].$$

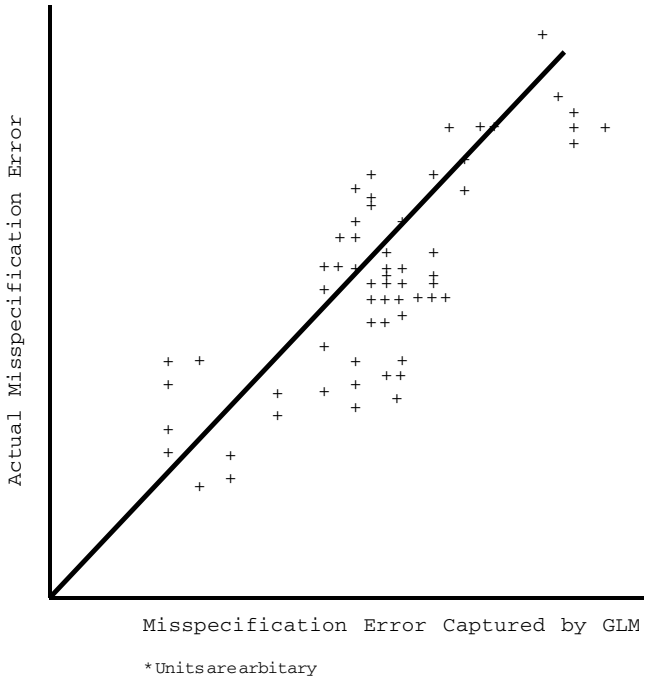
Research Design

We construct a simple synthetic example to test the performance of the three models. We assume two chains, and let chain 1 have retail locations at (1, 0) and (–1, 0). Chain 2 has two retail sites at (0, 0) and (0, 2). We assume (for simplicity) that only distance (representing, in effect, the variables present in the model) and misspecification error (representing the variables not present in the model) are predictive of choice. Thus,

$$(13)\quad U_{ij} = \exp[\alpha d_{ij} + ME_j(X_i) + \varepsilon_{ij}].$$

To operationalize equation (11), we set $\alpha = -2$, and $ME_2(X) = 0$ everywhere. We set $ME_1(X) = X + Y$, where X and Y are the two components of X, the geographical location. In other words, chain 1's misspecification error in-

Figure 2
RECOVERY OF GEOGRAPHICALLY LOCALIZED MISSPECIFICATION ERROR



creases to the Northeast (if X is North and Y is East). Also we will let δ_{ij} be distributed normally with mean 0 and variance 1. The central assumptions underlying this example are largely consistent (except for the error term) with those of the proposed model; therefore, we would expect the proposed geographically localized misspecification model (GLMM) to do well, if it is capable of doing well.

We let customers be sampled from a bivariate normal distribution with centroid (0, 0) and independent components of variance 1. We first estimate each of the three competing models, SL, UMM, and GLMM, on data for the four sites described previously. Then we introduce a new site, (–1, –1), for chain 2, and let each model predict the new market shares for chain 1 versus chain 2.

We repeat the above analysis 30 times and then collect error statistics on the ability to predict market share.

Results

It is of considerable importance whether the model misspecification can be captured by the GLMM model. Figure 2 shows graphically the relationship between $\widehat{ME}$ (X_i) and $ME(X_i)$ for each data point i, for replication one. We see that there is a strong positive correlation (.62) between the two and that it is highly significant ($p < .001$). The other replications produce roughly similar results. This indicates that the GLMM model is capable of modeling at least simple geographical variation in misspecification usefully.

Table 1 shows the ability of the three competing models to recover existing market shares and to predict market shares after the addition of the new store in chain 2. The actual market shares were closely approximated by simulating 10,000 points representing the population. We see that the simple logit (SL) model has difficulty recovering the existing market shares in the presence of geographically localized chain equity (mean SSE = 93.11).

Table 1
MARKET SHARE REPRODUCTION AND PREDICTION: SIMULATIONS

Reproduction of Estimated Market Shares

		Average Estimated Shares by Competing Models		
Chain	*Actual Shares*	*SL*	*UMM*	*GLMM*
1	46.69	39.91	48.93	50.09
2	53.31	60.09	51.07	49.91
Mean SSE		93.11	10.75	23.61

Prediction of Validation Market Shares

		Average Estimated Shares by Competing Models		
Chain	*Actual Shares*	*SL*	*UMM*	*GLMM*
1	45.26	37.99	50.04	48.69
2	54.74	62.01	49.96	52.32
Mean SSE		106.12	46.13	27.69

The GLMM model does better than SL (mean SSE = 23.61), but the UMM model recovers existing market share much more accurately than either of the two other models (mean SSE = 10.75). However, this apparent ability of the UMM model to fit the data is misleading. Simply fitting the estimation data is not a true test of the models' capabilities, because such fitting is based on the data itself, and therefore it is post hoc. One must also test the models' relative abilities to predict the market shares that will occur after a change in market configuration. We do this subsequently.

After the addition of chain 2's new site, the predicted market shares are less accurate for each model. This is not surprising, because in the Before case the choice data were used to calibrate the models. Here, the GLMM model performs best (mean SSE = 27.69), followed by UMM (mean SSE = 46.13) and SL (mean SSE = 106.12).

The market share prediction error for the GLMM model is about 40% less than the prediction error for UMM. This is an indication that the GLMM model can perform much better than the simple logit model or the uniform chain equity model in predicting market shares after the introduction of a new retail site, if model misspecification error is correlated with geographic location.

Statistical hypothesis tests comparing the mean SSE's (using t tests) confirmed that GLMM predicts significantly better (.05 level) than UMM and that GLMM and UMM models predict much better than simple logit.

In this cross-validation that uses holdout samples, no unfair inherent advantage is gained by any of the competing models, regardless of model complexity. Hence, the SSEs can be compared as is, without any adjustments for model parsimony. Therefore, we are confident in stating that the proposed GLMM model works best in this simulated exam-

Table 2
STORE LOCATIONS FOR EMPIRICAL STUDY

Chain	*Store*	*Location*	*Before*	*After*	*Comment*
1	1	(4.3, 4.2)	X	X	
	2	(3.1, 3.9)	X		closed
	3	(3.1, 3.1)		X	opened
2	1	(2.9, 1.2)	X	X	
	2	(3.2, 4.0)	X	X	
	3	(4.9, 2.9)	X	X	
3	1	(4.5, 4.0)	X	X	
4	1	(3.1, 3.5)	X	X	

ple in which geographically localized misspecification errors actually existed.

VALIDATION BY EMPIRICAL TESTING

The preceding section established that the GLMM model performed well in a simulated test in which the environment was largely consistent with the assumptions of the model. This provided internal validation of the potential usefulness of the GLMM model in predicting market shares after extension of a retail chain. We now need to establish evidence for the external validity of the model by testing its ability to estimate empirical market shares on the basis of actual changes in the configuration of a real market.

Research Design

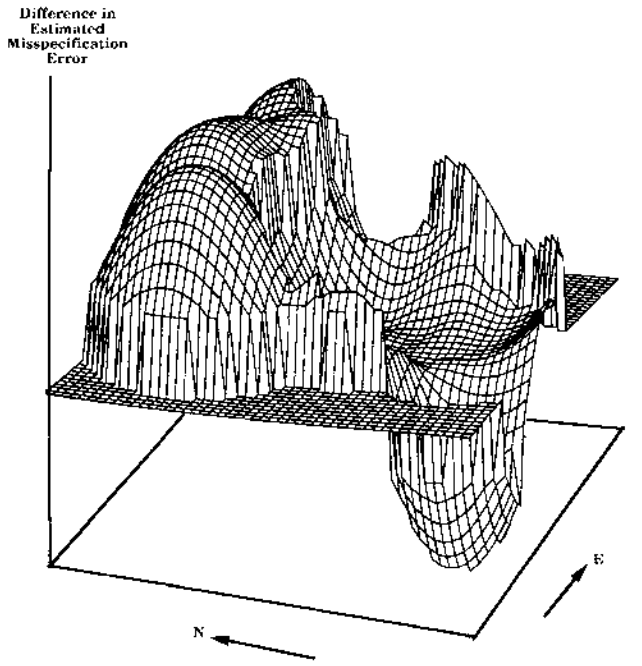
We examined a market in a small city that was dominated by four competing national grocery store chains. The locations of each chain's stores are given in Table 2. We drew a systematic random sample of 123 households from a telephone directory. The person in the household who does the most grocery shopping was asked to provide information about the store shopped at most frequently, including his or her perception of the price, service, variety of goods, and quality of goods of the grocery stores on a five-point scale. The respondent's geographic location was recorded on the basis of the listed address. This first wave of questioning provided the data necessary to estimate each of the three models.

Subsequent to the collection of the Before data, chain 1 added a new site and dropped one of its old sites (see Table 2). This provided a natural experiment in which the abilities of the three competing models to predict market share could be compared.

Following the reconfiguration of the market, we waited 13 weeks for the market share to settle down and then collected a new sample of 132 respondents, of whom we again asked for the store they shopped at most frequently. Thus, we were able to generate true market shares on the basis of the After sample, and we estimated market shares for each of the three competing models on the basis of the model coefficients obtained from the Before sample applied to the new market configuration.

Here, we used three specifications for the model: (M1), a simple gravity model with distance as the only independent variable; (M2), a more complete model with five predictor variables (price, service, variety, quality, and distance), but we included only the two significant predictor variables (for our data, of the five predictor variables, distance and price were the only two statistically significant variables); (M3), a fully specified model with all five predictor variables on which we had collected data from respondents in the first wave of data collection.

Figure 3
ESTIMATED GEOGRAPHICALLY LOCALIZED MISSPECIFICATION ERROR: DIFFERENCE BETWEEN CHAIN 3 AND CHAIN 4



Results

It is possible to examine the estimated geographically localized misspecification error graphically, two chains at a time. This can result in some striking observations. Figure 3 shows a three-dimensional representation of the relationship between chain 3's estimated misspecification error and that of chain 4 for the M1 specification.

The height of the plot reflects the difference between chain 3's estimated misspecification error and that of chain 4. Thus, a peak implies that chain 3 has a relatively positive (geographically localized) misspecification error compared to chain 4, whereas a valley indicates the reverse. One explanation for a peak may be that omitted variables favorable to chain 3 are systematically larger than average in that region.

It is clear from the plot that chain 3's estimated misspecification error increases with respect to chain 4's estimated misspecification error as we go farther on the Y dimension. We did further checking, based on the other questions in the survey, as to what could have caused this phenomenon. We divided the respondents into two groups, with group 1 located where chain 3's estimated misspecification error is larger than chain 4's and group 2 where chain 4's estimated

Table 3
MARKET SHARE REPRODUCTION AND PREDICTION: EMPIRICAL EXAMPLE

Reproduction of "Before" Market Shares										
		Estimated Shares by Competing Models								
Chain	*Actual Shares*	*SL*			*UMM*			*GLMM*		
		M1	*M2*	*M3*	*M1*	*M2*	*M3*	*M1*	*M2*	*M3*
1	62.4	29.2	32.5	33.1	66.2	65.8	65.6	72.2	70.5	70.1
2	14.4	32.5	30.5	29.8	17.0	16.5	16.5	12.2	12.8	12.9
3	15.6	14.1	14.2	14.8	12.3	12.8	13.1	11.5	12.2	12.3
4	7.6	24.3	22.8	22.2	4.5	4.9	4.9	4.0	4.5	4.6
SSE		1847.8	1509.8	1430.7	30.5	21.5	19.4	95.5	60.9	54.6
Prediction of "After" Market Shares										
		Estimated Shares by Competing Models								
Chain	*Actual Shares*	*SL*			*UMM*			*GLMM*		
		M1	*M2*	*M3*	*M1*	*M2*	*M3*	*M1*	*M2*	*M3*
1	80.8	31.9	34.6	35.4	66.1	68.4	68.9	87.1	86.0	85.8
2	10.6	33.1	32.0	31.6	17.0	16.1	15.1	8.3	8.8	8.9
3	5.3	13.3	12.0	11.8	12.4	11.5	11.3	4.0	4.5	4.5
4	3.3	23.7	21.5	21.0	4.5	4.0	4.5	.6	1.0	1.1
SSE		3577.2	2968.5	2857.7	308.9	222.9	199.3	54.0	36.21	33.4

Notes: M1 = Model with distance as the only predictor variable; M2 = Model with the two significant predictor variables (distance and price); and M3 = Model with all five predictor variables (distance, price, variety, service, and quality).

misspecification error is larger. We found several significant differences (at the .05 level) between group 1 and group 2. For example, group 1 spends more per week than group 2, and group 2 places more importance on price.

The implication is that group 1 is wealthier and that presumably chain 3's more upscale image and product line result in its increased draw for Group 1. What is interesting is that we were able to model chain 3's geographically localized appeal with respect to chain 4 *without having incorporated the variables that caused the difference in the model*.

The GLMM model was able to respond to these effects without explicitly being specified to include them, because of the interaction between these variables and geographic location. The UMM model would not have been able to respond to these effects unless the model was fully specified. This basically shows that the proposed GLMM model accounts for model misspecification.

As Table 3 shows, both the GLMM and UMM models were able to reproduce the market shares for the old market configuration, whereas the simple logit model was unable to accommodate the strength of chain 1 and therefore provided a very poor market share fit. Here, each model provided three market share estimates (M1 = Model with distance as the only predictor variable; M2 = Model with the two significant predictor variables [distance and price]; M3 = Model with all five predictor variables [distance, price, variety, service, and quality]).

As one would expect, the sum of square error (SSE) was lowest for the model with all five predictor variables (M3). However, there is not much improvement in the predictive power for SL, UMM, and GLMM as we increase the model specification from M1 (or M2) to M3. However, the true test of these models is whether they are able to predict market shares after the configuration changes.

For the new configuration, we see in Table 3 that the GLMM model performed better in predicting the new market shares. GLMM's sum of square error (54.0, 36.21, and 33.4 for M1, M2, and M3, respectively) was about one sixth that of UMM (308.9, 222.9, and 199.3 for M1, M2, and M3, respectively), and both GLMM and UMM far outperformed the SL model (3577.2, 2968.5, and 2857.7 for M1, M2, and M3, respectively).

No statistical hypothesis testing was possible, because of the sample size of one (an actual natural experiment), but the results suggest that modeling geographically localized model misspecification may be useful. In this cross-validation, using holdout samples, no unfair inherent advantage is gained by any of the competing models, regardless of model complexity. Hence, the SSEs may be compared as is without any adjustments for model parsimony, and we conclude unambiguously that GLMM's predictive performance is superior in this test.

Interestingly, the GLMM approach improved the prediction error by about the same percentage, regardless of the completeness of the predictor model. This illustrates the potential usefulness of our approach, even in cases where the prediction model is well specified.

CONCLUSIONS

Combining logit choice models with nonparametric density estimation, we provide a statistical methodology for approximating the geographically localized misspecification error (residuals of unobserved variables) in retail store choice models and incorporating the concept in models of store choice.

Unlike consumer goods, retail chains have a geographic component. We can use this geographic information to our advantage, because it may interact with other variables. Be-

cause the geographic locations of store sites and customers can be easily and accurately measured, we can gain knowledge of the market by considering how residuals of unobserved variables in logit models of store choice vary geographically.

Our experiments, which are based on both simulated and empirical data, suggest that incorporating geographically localized model misspecification may greatly improve the ability to model and predict market shares.

Our method is also capable of addressing the problem of spatial nonstationarity of the parameters by reexpressing that problem as an omitted variables problem and then estimating the localized effect of the omitted variables.

The geographically localized misspecification error can be viewed as the advantage of one chain over another in attracting customers. Hence, another way of thinking about this method of capturing misspecification error of the retail store choice model is that it captures the equity of retail chains (chain equity). Given that the omitted (unobserved) variables interact with geographic location and that they often tend to be differences such as image and reputation, the misspecification error may correspond with the concept of brand equity, as it is usually discussed.

We used a very simple choice model to illustrate the concept of geographically localized model misspecification error. However, it should be possible to extend this approach to more sophisticated retail attraction models, such as a revised central place model (Ghosh 1986) or a hierarchical choice model (Gensch 1987). We leave such extensions for further research.

To summarize, retail store choice models are often misspecified. The unobserved variables in these models can be correlated with geographic location, permitting the misspecification errors to be geographically localized. Here, we propose an approach to capture these geographically localized residuals by combining nonparametric density estimation with logit choice models. Our validation tests indicate that modeling geographically localized residuals can be useful in predicting consumer choice and retail chain market shares.

REFERENCES

Achabal, Dale D., W. L. Gorr, and Vijay Mahajan (1982), "MULTILOC: A Multiple Store Location Decision Model," *Journal of Retailing*, 58 (Summer), 5–25.

Batsell, Richard R. and Leonard M. Lodish (1981), "A Model and Measurement Methodology for Predicting Individual Consumer Choice," *Journal of Marketing Research*, 18 (February), 1–12.

———— and John C. Polking (1982), "A Generalized Model of Market Share," *Marketing Science*, 4 (Summer), 177–98.

Bucklin, Louis P. (1971), "Retail Gravity Models and Consumer Choice: A Theoretical and Empirical Critique," *Economic Geography*, 47 (October), 489–97.

Craig, C. Samuel, Avijit Ghosh, and Sara McLafferty (1984), "Models of the Retail Location Process: A Review," *Journal of Retailing*, 60 (Spring), 5–36.

Currim, Imran S. (1982), "Predictive Testing of Consumer Choice Models Not Subject to Independence of Irrelevant Alternatives," *Journal of Marketing Research*, 19 (May), 208–22.

Donthu, Naveen and Roland T. Rust (1989), "Estimating Geographic Customer Densities Using Kernel Density Estimation," *Marketing Science*, 8 (Spring), 191–203.

———— (1991), "Comparing Market Areas Using Kernel Density Estimation'' *Journal of the Academy of Marketing Science*, 19 (4), 323–32.

Fotheringham, A. Stewart (1980), "Spatial Structure and the Parameters of Spatial Interaction Models," *Geographical Analysis*, 12, 33–46.

———— (1981), "Spatial Structure and Distance-Decay Parameters," *Annals of the Association of American Geographers*, 71 (September), 425–36.

———— (1985), "Spatial Competition and Agglomeration in Urban Modeling," *Environment and Planning A*, 17, 213–30.

Gensch, Dennis H. (1987), "A Two-Stage Dissaggregate Attribute Choice Model," *Marketing Science*, 6 (Summer), 223–39.

———— and W. W. Recker (1979), "The Multinomial, Multiattribute Logit Choice Model," *Journal of Marketing Research*, 16 (February) 124–32.

Ghosh, Avijit (1986), "The Value of a Mall and Other Insights From a Revised Central Place Model," *Journal of Retailing*, 62 (Spring), 79–97.

———— and C. Samuel Craig (1983), "Formulating Retail Location Strategy in a Changing Environment," *Journal of Marketing*, 47 (Summer), 56–66.

———— (1984), "Parameter Nonstationarity in Retail Choice Models," *Journal of Business Research*, 12 (4), 425–36.

Guadagni, Peter M. and John D. C. Little (1983), "A Logit Model of Brand Choice Calibrated on Scanner Data," *Marketing Science*, 2 (Summer), 203–38.

Huff, David L. (1964), "Defining and Estimating a Trade Area," *Journal of Marketing*, 28, 34–30.

Inagaki, Nobu (1977), "Two Errors in Statistical Model Fitting," *Annals of the Institute of Statistical Mathematics*, Part A, 29, 131–52.

Kamakura, Wagner A. and Rajendra K. Srivastava (1984), "Predicting Choice Shares Under Conditions of Brand Interdependence," *Journal of Marketing Research*, 21 (November), 420–34.

Louviere, Jordan (1984), "Using Discrete Choice Experiments and Multinomial Logit Choice Models to Forecast Trial in a Competitive Retail Environment: A Fast Food Restaurant Illustration," *Journal of Retailing*, 60 (Winter), 81–107.

Luce, R. Duncan (1959), *Individual Choice Behavior*. New York: John Wiley & Sons, Inc.

McFadden, Daniel (1980), "Economic Models for Probabilistic Choice Among Products," *Journal of Business*, 53 (3), 513–29.

Rust, Roland T. and Julia A. N. Brown (1986), "Estimation and Comparison of Market Area Densities," *Journal of Retailing*, 62 (Winter), 410–30.

———— and David C. Schmittlein (1985), "A Bayesian Cross-Validated Likelihood Method for Comparing Alternative Specifications of Quantitative Models," *Marketing Science*, 4 (Winter), 20–40.

SAS Institute Inc. (1987), *SAS/GRAPH Guide for Personal Computers; Version 6 Edition.* Gary, NC: SAS Institute Inc.

Silverman, B. W. (1986), *Density Estimation for Data Analysis and Statistics*. New York: Chapman and Hall.

Swait, Jeffrey and Jordan Louviere (1993), "The Role of the Scale Parameter in the Estimation and Comparison of Multinational Logit Models," *Journal of Marketing Research*, 30 (August), 305–14.