Chapter 6

Integrating a Hierarchical Bayes Gravity-like Model into a Retail Chain's IT System

Thomas C. Eagle

6.1 Introduction

My professional career after graduating as one of Dr. Gerard Rushton's doctoral students in Geography took a path into marketing research and consulting where I became an 'engineer' instead of an academic. My key job was and has been to find solutions to complex marketing questions that almost always involve the modeling of consumer behavior. This over time has demanded the solution of more and more complex problems with large data sets and systems of models in order to capture the nuances of behaviors that clients wish to predict.

Also as time has gone by, more clients have needed a Decision Support System (DSS) in order to understand the predictions of our analyses. These systems were and still are theoretically inelegant or imperfect. Engineers in the business world tend to solve extremely complex problems in a short time frame, frequently with poor data, and in the face of client constraints. Prediction often outweighs explanation. From my perspective, one of Rushton's greatest influences was to shape the way I could tackle and develop systems of models that described or predicted these data [1].

These subsequently for me have included systems for modelling the impacts on company profits of pricing and

availability of menu items in fast food restaurants, sandwich chains, and full service restaurants; for developing optimization systems to maximize profits for major airlines; for customizing the market segmentation of the research division of a major energy provider in terms of specific strategic and tactical requirements; and for integrating a hierarchical Bayes gravity-like model into the IT system of a major national retail chain.

We do not explain the theory and procedure of hierarchical Bayes modeling used in this representative DSS for integrating a hierarchical Bayes gravity-like model into the IT system of a major national retail chain. Nor can we show the details of the final system of models as this is proprietary to the client. We therefore discuss the general details of the modeling framework and the data included in, and excluded from, the calibrated model; how it was integrated into the client's IT system; and how it is now being used firm-wide.

6.2 Introduction to one of my Company's DSS

This client's posed problem is indeed representative for me: A major national retail sales chain required a better process and statistical model to predict the viability of potential new store locations. Members of the chain's site evaluation team previously examined sites by hand. They had access to an aggregate distance decay model for predicting shares of sales in block groups around new sites. Their model however did not incorporate sales potential, the impact of competition, the impact of sister stores in the same spatial area, or the characteristics of the new market area itself on the predictions. No updating of the model or its predictions was available. The aggregate distance decay model was

incorporated into a GIS, but it was a standalone model used only by the site evaluation team.

We proposed to improve this system in at least three ways by: (1) Improving the predictions of the model at the block group level for aggregate average weekly sales in stores, while incorporating competitive and existing sister store impacts. (2) Integrating the modeling system into the chain's IT system to enable both seamless use of GIS and storage of data internally in the chain's IT system, together with updates under their control. (3) Developing a system that would allow the site evaluation team to calculate new predictions for others to use in a controlled way if they inserted new stores, competition, and altered parameters.

6.3 The Modeling System

Hierarchical Bayes (HB) models were used to comprise a store site sales evaluation system. Hierarchical Bayes models are a form of multi-level random effects models, and their conceptual structure is illustrated in Figure 6.1 (e.g., [2, 3]). There is an upper-level model governing the heterogeneity across the lower-level model. The lower-level model is a gravity-like model predicting share of sales from each block group comprising the market area. One result of using the hierarchical Bayes model is that each existing store has its own set of estimated gravity-like model parameters across the posterior draws of the Monte Carlo Markov Chain simulations.

The upper-level model therefore governs and predicts the lower-level model parameters. This upper-level model was a multivariate regression of the lower-level gravity-like model parameters on the set of store and market area specific characteristics. This upper-level multivariate regression model may consequently produce predictions of lower-level

model store parameters for a new site with assumed store and market area characteristics. The first statistical objective is therefore to produce a reasonable upper-level model to predict a store's lower-level gravity-like model parameters. Then, the second statistical objective is to predict a set of new site-specific gravity-like model parameters that would improve upon those of the original aggregate gravity-like model used by the chain.

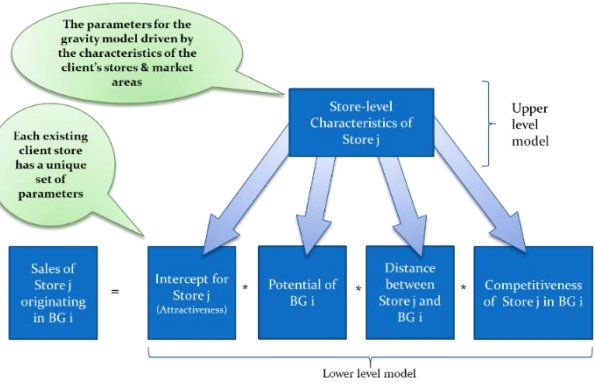

Figure 6.1 The basic Hierarchical Bayes model. (Color copy online.)

Existing store sales and market area data were used as data for calibrating the model. In the HB framework, each store has parameters estimated for the variables used to represent the lower-level model components. This model predicts the sales originating from each block group. Each store's market area and store-level characteristics are used in the upper-level of the HB modeling to predict the values of its lower-level gravity-like parameters. The final estimated model is essentially a zero-inflated binomial model of the proportion of block group sales captured by the chain's stores.

There however were particular data constraints for this modeling. For example, the sales data originating from each

block group was incomplete, so that there was an unknown amount of spatially-incomplete information in the dependent variable of our model. A system of hierarchical Bayes models was constructed to account for this uncertainty in the dependent variable. The system ultimately consisted of a model to predict the likelihood of sales originating for each block group, followed by a model predicting the share of sales from each block group inferred from the likelihood of sales originating there. The same set of upper- and lower-level model predictors were used in both models.

6.4 Data

The market area for each store was designated as the 10-mile radius around each one. Within that 10-mile radius we had 1-mile, 3-mile, 5-mile, and 5 to 10-mile aggregated zonal census and food sales data (other zonal boundaries were examined as well). These data were subjected to a principal components analysis to reduce the number of variables and thus remove their multicollinearity. Multivariate regressions in the upper level model also included independent variables for each store's unique characteristics, including gross square footage, retail sales footage, and store experts' qualitative and quantitative measures of the site's characteristics. These latter measures included ease of access to the store and parking; location relative to major thoroughfares, the CBD, trains and subways; and information about the locations and strengths, or potentials of sister stores and the store's competition.

Last, data for the surrogate of the weekly sales data originating from each block group were incomplete and did not aggregate up to the average weekly total sales in stores. We had many block groups with zero sales when, in fact, sales likely did originate from those block groups. As a result, a zero-inflated hierarchical Bayes model system was

developed. One model component predicts the likelihood of any sales originating from a block group; and then a second component predicts the share of sales from the block group based upon the likelihood of any sales from the block group.

Otherwise, block group potential data included all the available census information, sales data for various categories of food purchases, total sales potential (total dollars spent on food items in the block group), and its location. A principal components analysis was conducted similarly to the upper-level model using these data across all block groups in the US. The resulting factor scores were measures of block group potential.

Straight-line distance measured in miles from each block group to each store was used in the models. Road distances and travel times that were either directly measured or computed from GIS system capabilities were not used at the direct request of the client. Our model as a result does not capture the direct impact of any spatial barriers that may influence the shoppers' behaviors.

One of four indices of the impact of competition on the sales originating from each block group was a distance-weighted index for the strength of each competitive store within a 10-mile radius of each block group. In addition, a distance-weighted intervening opportunity index was calculated in an analogous manner. These two indices were also computed for the client's sister stores within the 10-mile radius of each block group. These were adapted from the original ones of another author for measuring patient flows using Bayesian methods [4]. Zonal summary measures of these competitive indices were used in the upper-level model.

6.5 Estimation and Final System of Models

The models were calibrated with available Markov Chain Monte Carlo (MCMC) routines that utilized a Gibbs Sampler to predict the binomial likelihood of (1) sales originating from each block group; and (2) the proportion of a block group's total sales potential going to the client's store based upon the likelihood of sales originating from that block group. These models were validated using a hold-out sample of existing stores.

This series of calibrated models were further refined for sales' predictions in stores under competition from nearby sister stores and competitive stores. The final system of models predicted average weekly stores sales both with good statistical explanation (R-squared of 65%), and in comparison with actual average weekly sales (Figure 6.2). Note that the aggregate analogous gravity-like model had a much lower R-squared of 26% explained variation.

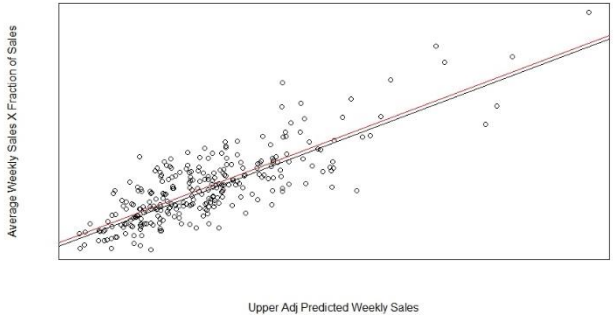

Figure 6.2 Predicted versus actual average weekly store sales from the modeling system. Note the upper (red) solid line is the simple regression of predicted sales on actual sales, and the lower solid line is the diagonal. (Color copy online.)

6.6 Integration of the Modeling System into Retail Chain's IT system

In the end, the lack of a centralized IT system across the units of the client's organization presented a unique system's development issue. These units included marketing, site planning, advertising, and sales – and the modeling system obtained data from some of these units and IT. Channels allowing access to these data were required. An integrated data access system was constructed for the modeling system's acquisition of required data with minimal user input. This unified the different data contained across these systems, the coordination with the client's IT group, and the knowledge of required inputs for the modeling system. Data were in some cases transferred to a centralized system; while in other cases direct access channels were developed.

Model results were also networked within the chain's IT system. The modeling system was designed for a user to simulate new and existing store sites and to view results. The user has the option of saving the results locally. Simulations thought to be useful for storing in the chain-wide data base are submitted to a modeling system administrator for him or her to evaluate and to approve a request for storage of results in the chain-wide IT system. Saved results and inputs may be called from the IT system, but not modified without an administrator's approval. Of course, the chain's IT system administrators were heavily involved in the design and implementation of the data access and storage system (Figure 6.3).

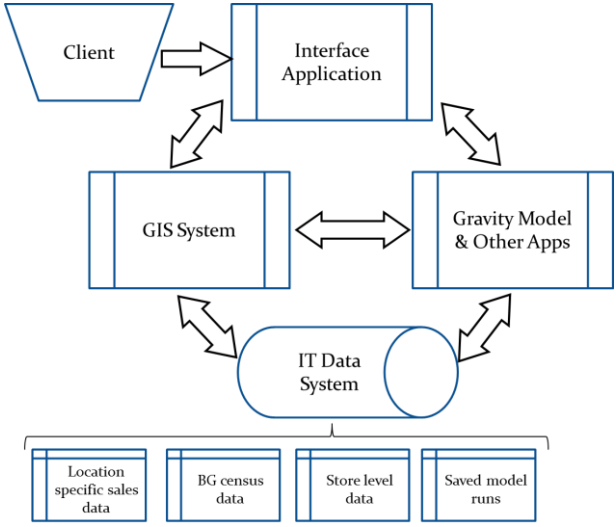

Figure 6.3 Representation of the implemented IT system. (Color copy online.)

The user interface was implemented within the GIS system using the GIS system's programming language. Some modules were written for executing routines outside of the GIS system. These modules ran the modeling system and saved results appropriately. The basic modeling system is run from a designed user interface that permits a user without GIS-knowledge to evaluate a site. Evaluation of a new site requires additional inputs including the longitude and latitude of the new site; the site's required physical and evaluative characteristics; and the type of tabular and/or mapping reporting desired by the user.

The user interface when reloading an existing store's site and model results, promotes 'what-if' scenarios from changing the existing store's characteristics; changing existing sister and competitive store characteristic; and inserting new competitive and sister stores into the market area and evaluating the impact of these new stores on existing

stores. Furthermore, an expert user with knowledge of GIS may adjust predictions down to the block group level, modify the aggregate store predictions, and alter model system inputs to refine the modeling system's results. An expert can thus refine the new or existing store's results using first-hand knowledge not included in the modeling system.

6.7 Conclusion

The modeling system has now been fully integrated into the firm-wide IT system, for example, with the results that: (1) The process of evaluating new sites has been streamlined. (2) There have been more accurate predictions of store sales at new sites opened since the modeling system was implemented. (3) Various units across the retail chain not directly linked to site evaluation are using the system to improve marketing strategy, customize store inventory, and evaluate new store concepts.

Finally, the modelling system is re-estimated and refined on an annual basis. Furthermore, the retail chain continues to request model refinements for the improvement of outlier predictions, and new applications beyond its original scope. Such requests are a clear sign of the integration of a complex modeling system into a company-wide environment. The modeling system's developers are always striving to improve its predictions as well as expanding its capabilities to address more specific and refined requests.

References

[1] G. Rushton, "The Roepke Lecture in Economic Geography: Location theory, location-allocation models, and service development planning in the Third World," Economic Geography, vol. 54, no. 2, pp. 97-120, 1988.

[2] P. D. Congdon, Applied Bayesian Hierarchical Methods. London, UK: CRC Press, 2010.

[3] J. LeSage and R. K. Pace, Introduction to Spatial Econometrics. London, UK: CRC Press, 2009.

[4] P. D. Congdon, "A Bayesian approach to prediction using the gravity model, with an application to patient flow modelling," Geographical Analysis, vol. 32, no. 3, pp. 183-197, 2000.