On combining revealed and stated preferences to forecast customer behavior: three case studies

Peter C. Verhoef[1]

Erasmus University Rotterdam

Department of Marketing and Organization

Philip Hans Franses

Erasmus University Rotterdam

Econometric Institute

Department of Marketing and Organization

Abstract

Many companies collect stated preference data (SP) like intentions and satisfaction as well as revealed preference data (RP) like actual purchasing behavior. It seems relevant to examine the predictive usefulness of this information for future revealed preferences, that is, customer behavior. In this paper we address this issue by considering three case studies.

Key words: Customer behavior; forecasting; revealed preferences; stated preferences

---

[1] Address for correspondence: Peter C. Verhoef, Erasmus University Rotterdam, School of Economics, Office H15-12; P.O. Box 1738, NL-3000 DR Rotterdam, The Netherlands, Tel. +31 10 408 2809; Fax. +31 10 408 9169, e-mail: verhoef@few.eur.nl

1. Introduction

At present, many companies collect stated preference data (SP) among large samples of customers. Examples are annual surveys concerning, for example, stated loyalty, satisfaction and future needs. At the same time, many companies also keep track of actual behavior of their customers by collecting revealed preference data (RP) in their customer databases. Examples of these data are records on past purchases, additional purchasing behavior, and number of contacts with the company. These data sources provide the manager of the company with important metrics, such as brand preference, satisfaction and customer retention (Kaplan and Nortan, 1992). As the collection of SP and RP data is costly, natural questions to ask concern the predictive usefulness of current SP data for future revealed preferences and the potential relevance of past behavior (measured by the RP data) for predicting future behavior. In this paper we aim to address these questions jointly by considering three case studies.

Forecasting marketing metrics, such as sales, market shares and customer retention, is an important activity of marketing researchers in large companies. Market share and sales forecasting is usually done using time series models (see for example Hansens, Parsons and Schultz, 2001; Fok and Franses, 2001; Kumar, 1994; Wittink et al., 1988), in which information on past metrics is used to predict the future value of these metrics. Besides the use of these past metrics, market research firms often use self-revealed preferences or purchase intents to forecast sales, see Leeflang and Wittink (2000). Purchase intents are especially used to forecast sales of new products, as for these products no data on past sales are available (Bemmaor, 1995; Louviere, Henscher and Swait, 2000). An important disadvantage of purchase intents is its potentially limited predictive validity, as individual-level observed behavior can differ from planned behavior (Bemmaor, 1995; Morwitz, Steckel and Gupta, 1997).

An important development within marketing research is that companies store and use data on the purchase behavior of their customers in databases. This is often referred to as database marketing, which in turn can be used for customer relationship management (Verhoef and Hoekstra, 1999; Winer, 2001). Recently, a number of studies have developed models that used these revealed preference data as well as stated preference data from a survey to describe customer behavior (Bolton, 1998; Bolton and Lemon, 1999; Mittal and Kamakura, 2001; Verhoef, Franses and

Hoekstra, 2001). In general these models show significant in-sample effects of the self-reported data on metrics such as relationship duration, service usage and cross-buying. It is of course also interesting to examine how this in-sample fit carries over to out-of-sample forecasting. We are aware of only a single study, which addresses a similar issue. Van der Poel, DeKimpe and Leunis (2000) find not much additional predictive value of SP data in a direct mail response model. In this paper we aim to contribute to the knowledge base by studying the predictive power of SP variables, of RP variables and of both, jointly. For this purpose we consider three case studies, all concerning one company. This may seem rather restrictive, but perhaps it is important to stress here that combined RP and SP data are not easy to collect and that access to available databases is limited (due to confidentiality).

The structure of this paper is as follows. In Section 2 we continue with a brief discussion on why combining RP and SP data might be better than just looking at RP or SP data only. The three case studies are presented in Section 3. We end with a conclusion in Section 4.

2. Forecasting with SP and RP data

In this section we first discuss some reasons why SP data can be of interest in marketing. Next, we briefly discuss the strengths and weaknesses of RP and SP data. Finally, we provide reasons why the combination of data sources may lead to better out-of-sample forecasting results.

Typically there is skepticism about relying on what consumers say they will do, which is based on comparing their statements with what they actually do. Indeed, there is often a substantial discrepancy between stated preferences and future behavior (Morwitz, Steckel and Gupta, 1997). Despite this discrepancy, there are a number of compelling reasons when one should be interested in SP data (Louviere, Henscher and Swatt, 2000, p. 21). First, the demand for new or improved products with new attributes has to be estimated (Wittink, 2001). Second, explanatory variables such as prices have little variability in the market place and/or are highly collinear, and hence might have little predictive value. Third, new variables can be introduced that may explain choices which are not captured by RP data, like satisfaction. Fourth, observational data can be expensive to collect.

Due to an increasing use of customer databases, many more firms are able to collect RP data from their customers at a reasonable cost. Hence, some of the reasons for using SP data only have become obsolete. The next question is then if collecting SP data, additional to RP data, is worth the trouble at all? In order to understand the added value of SP data in these contexts, we need to have a closer look at the features of SP and RP data. RP data have a high reliability and validity, which is usually due to the measurement system. An additional important characteristic of RP data is that it concerns the world as it is now. As a result, these data only describe currently existing choice options and these data are limited to market and personal constraints. Due to these characteristics, these data are perhaps pretty well suited for short-term forecasting. However at the same time, these characteristics make RP data less reliable for forecasting over a longer range. In contrast, SP data can contain individual-specific information about the future. For example, although customers have been customers for a long time, recent experiences may have established negative evaluations of the company, in turn resulting in high intentions to switch. These intentions are not captured in the customers' observed behavior, at least not exactly. Moreover, SP data are not limited to the current market conditions and personal constraints. An important disadvantage of SP data, though, is their low reliability and validity, but this might be improved by using the proper measures (Juster, 1966; Morwitz, Steckel and Gupta, 1997). As a consequence of these characteristics, SP data might be viewed as especially appropriate to predict structural changes that occur over longer time periods (Louviere, Henscher and Swait, 2000).

In sum, SP and RP data are seen to each have their own strengths and weaknesses. Traditionally, these two data sources have been considered separately in (marketing) forecasting research. A stream of research has focused on developing time series type of models to forecast metrics, while other researchers have focused on the forecast accuracy of SP data using conjoint models or other approaches. A few recent studies advocate to combine both data sources (Bolton, 1998; Verhoef, Franses and Hoekstra, 2001). In particular, Louviere, Henscher and Swait (2000; p. 231) state that the strengths of both data sources could be exploited and weaknesses ameliorated by pooling both data sources. This process is also called data enrichment. This enrichment should provide more robust parameter estimates and should increase confidence and accuracy in predictions. To see whether this is the case, we examine three case studies in the next section.

3. Case studies

We have access to the database, containing RP and SP data, from a large Netherlands-based financial service provider. We have data of a representative sample of approximately 2000 customers, concerning two time periods T_0 and T_1. RP data are available from the customer database and include variables, such as the type of service purchased, the number of services purchased and relationship age. SP data were collected among these customers at T_0. These SP variables concern summated questions on satisfaction, commitment and word-of-mouth communication. The latter variable is often used as a measure for purchase intention (see Mittal, Kumar and Tsiros, 1999). An overview of the included RP and SP data in our empirical models is provided in Table 1.

The variables to be explained in our case studies concern customer behavior between T_0 and T_1, measured at T_1. The period between T_0 and T_1 covers a year. Each study has a different variable to be explained. First, we we aim to predict customer retention. Next, we consider cross-buying, which is the difference in the number of services purchased at the supplier between T_0 and T_1. Finally, customer share (the ratio of services purchased by a customer at asupplier over the total purchases in a service category of that customer) is considered in the third case study. In all three cases we use approximately (a randomly selected) two-third of the available sample to estimate the model, while we use the remaining part of the sample as a holdout to assess the predictive performance of these models. Thereby, we compare the performance of the following three models, that is, a model with SP and RP variables, a model with only RP variables and a model with only SP variables.

It should be noted here that our forecasts concern the cross-sectional dimension, that is, we leave out individuals to evaluate the models. It would also have been interesting to have a model constructed for T_0 and T_1 data to forecast data at T_2. As of yet, we do not have such data, so we postpone this issue to further work.

Customer Retention

Customer retention is a binary variable, and therefore we use the logit model to forecast it. In order to compare the predictive performance of the three models, we

consider the log likelihood (LL) (Franses and Paap, 2001, p.59, equation (4.33)). The LL values of the three models for the estimation sample and the holdout sample are given in Table 2. The likelihood ratio (LR) test for the significance of the SP variables in the joint model is not significant at the 5% level, while the LR test for the RP variables is significant at the 1% level. For the holdout sample, we observe that the likelihood is highest for the joint model, and that only including SP data leads to very poor forecasts.

Cross-buying

The ordered logit model is used predict cross-buying, as in our database cross-buying can be considered as an ordinal variable taking only values –1, 0, 1 and 2. The expression of the relevant log likelihood is given in Franses and Paap (2001, p. 119, equation (6.20)). The LL values of the three models are displayed in Table 3. Again, the inclusion of SP variables in the joint model is not significant at the 5% level, while the inclusion of RP variables is significant even at the 0.1% level. In the holdout sample the difference in LL values between the full model and model with only RP variables is negligible. Again, the model with only SP data performs rather poorly.

Customer Share

Customer share can be considered as a continuous variable. We therefore rely on a standard regrression model with the log of customer share as the dependent variable (see Bowman and Narayandas, 2001). In this model we also include customer share at T_0, which can also be viewed as a RP variable. To compare the predictive performance of the three models we consider the Mean Absolute Percentage Errors (MAPE), see Table 4. Again, the model with only SP variables has the worst predictive performance. The difference in predictive power both in-sample and out-of-sample between the two other models is very small.

4. Conclusion

In this paper we examined whether the combination of RP and SP variables in forecasting models for customer behavior yields better results than for models with

only RP or SP variables. The relevant literature suggests that this combination could be fruitful (Louviere, Henscher and Swait, 2000). Our exploratory results are that in-sample there is just a modest gain from also including all SP variables, while out-of-sample forecasts tend to be equallly good or slightly better when these variables are included. Additionally, we find that models with only SP variables perform very poorly. Hence, the collection of both types of data seems useful, where in particular RP data are not to be missed as these have substantial forecasting power.

We used only three case studies. We therefore identify the following specific avenues for further research. First, in order to obtain more general insights, studies on the forecasting merits of SP and RP data could be done for other markets and other databases. Second, these studies may focus on the conditions, which enhance the combination of SP and RP to have more forecasting power. For example, in business markets, SP data obtained from purchase managers could provide much more forecastibility. This research would then be in line with prior research on the predictive validity of purchase intents (Morwitz, Steckel and Gupta, 1997). Finally, the models used in this paper did not account for customer heterogeneity. In future research, the forecasting performance of models considering such heterogeneity could be examined.

REFERENCES

Bemmaor, A.C. (1995), Predicting behavior from intention-to-buy measures: The parametric case, *Journal of Marketing Research*, 32 (2), 176-91

Bolton, R.N. (1998), A dynamic model of the duration of the customer's relationship with a continuous service provider: The role of satisfaction, *Marketing Science*, 17 (1), 45-65

Bolton R.N. & Lemon K.N. (1999), A dynamic model of customers' usage of services: Usage as an antecedent and consequence of satisfaction, *Journal of Marketing Research*, 36 (2), 171-86

Bowman D. & Narayandas D. (2001), Managing customer-initiated contacts with manufacturers: The impact on share of category requirements and word-of-mouth behavior, *Journal of Marketing Research*, 38 (3), 281-97

Fok D. & Franses P.H. (2001), Forecasting market shares from models for sales, *International Journal of Forecasting*, 17, 121-128

Franses P.H. & Paap R. (2001), *Quantitative Models in Marketing Research*, Cambridge: Cambridge University Press

Hansens D.M., Parsons L.T. & Schultz R.L. (2001), *Market Response Models: Econometric and Time Series Analysis*, Boston: Kluwer Academic Publishers

Henscher D., Louviere J. & Swait J. (1999), Combining sources of preference data, *Journal of Econometrics*, 89, 197-221

Juster F.T. (1966), Consumer buying intentions and purchase probability: an experiment in survey design, *Journal of the American Statistical Association*, 61, 658-696

Kaplan R.S & Nortan D.P. (1992), The balanced scorecard: measures that drive performance, *Harvard Business Review*, 70 (1), 71-79

Kumar V. (1994), Forecasting performance of market share models: an assessment, additional insights, and guidelines, *International Journal of Forecasting*, 10, 295-312

Louviere J.J., Henscher D.A. & J.D. Swait (2000), *Stated Choice Methods: Analysis and Application*, Cambridge: Cambridge University Press

Leeflang P.S.H. & Wittink D.R. (2000), Building models in marketing: past, present and future, *International Journal of Research in Marketing*, 18, 105-126

Mittal V, Kumar P & Tsiros M. (1999), Attribute-level performance, satisfaction, and behavioral intentions over time: A consumption-system approach, *Journal of Marketing*, 63 (2), 88-101

Mittal V. & Kamakura W.A. (2001), Satisfaction, repurchase intent, and repurchase behavior: Investigating the moderating effect of customer characteristics, *Journal of Marketing Research*, 38 (1), 131-142

Morwitz V.G., Steckel J.H. & Gupta A. (1997), When do purchase intentions predict sales?, *Working Paper Marketing Science Institute*, June, Report No. 97-112

Van der Poel D., DeKimpe M.G. & Leunis J.V. (2000), Segmentation bases for database marketing: Which variables matter?, *Proceedings Marketing Science Conference,* Los Angeles

Verhoef P.C., Franses P.H. & Hoekstra J.C. (2001), The impact of satisfaction and payment equity on cross buying: A dynamic model for a multi-service provider, *Journal of Retailing*, 77, 359-78

Verhoef P.C. & Hoekstra J.C. (1999), The status quo of database marketing in the Dutch FMCG-Industry, *Journal of Market Focused Management*, 3, 313-332

Winer, R.S. (2001), A framework for customer relationship management, *California Management Review*, 43 (4), 89-108

Wittink D.R. (2001), Forecasting with conjoint analysis, Principles of Forecasting: A Handbook for Researchers and Practitioners (ed. Armstrong J.S.), Boston: Kluwer, 147-167

Wittink D.R., Addona M., Hawkes W. & Porter J. (1988), SCAN*PRO: The estimation, validation and use of promotional effects based on scanner data, *Working Paper*, Johnson Graduate School of Management, Cornell University

Table 1: Explanatory variables

(for details on measurement, see Verhoef, Franses and Hoekstra, 2001)

Revealed Preferences (RP)	Stated Preferences (SP)
Dummy ownership of Co-insurance	**Satisfaction**
Dummy ownership Car-insurance	**Word-of-Mouth**
Dummy ownership Life-insurance	**Commitment**
Dummy ownership Damage-insurance	
Dummy ownership Travel-insurance	
Dummy ownership Bank Product (for example a loan)	
Total number of services purchased	
Relationship Age	
Member of Loyalty Program	
Number of Points in Loyalty Program	

Table 2: Log likelihood values of the binomial logit model for retention

Included Variables	Number	In Sample (n=1489)	Out of Sample (n=480)
RP + SP	13	-293.6035	-85.8312
RP	10	-297.0530	-88.7534
SP	3	-361.2225	-103.048

Table 3: Log likelihood value of the ordered logit model for cross-buying

Included Variables	Number	In Sample (n=1489)	Out of Sample (n=480)
RP + SP	13	-737.4280	-251.4627
RP	10	-740.9737	-251.2498
SP	3	-854.6957	-284.2137

Table 4: Mean absolute prediction errors in the regression model for customer share

(for a smaller sample due to data availablity)

Included Variables	Number	In Sample (n=553)	Out of Sample (n=185)
RP + SP	14	13.06	15.44
RP	11	12.99	15.42
SP	3	60.28	48.99