ADVANCES IN BUSINESS AND MANAGEMENT FORECASTING

VOLUME 6

KENNETH D. LAWRENCE
RONALD K. KLIMBERG
Editors

ADVANCES IN BUSINESS AND MANAGEMENT FORECASTING

ADVANCES IN BUSINESS AND MANAGEMENT FORECASTING

Series Editor: Kenneth D. Lawrence

Recent Volumes:

Volume 1: Advances in Business and Management Forecasting: Forecasting Sales

Volume 2: Advances in Business and Management Forecasting: Forecasting

Volume 3: Advances in Business and Management Forecasting

Volume 4: Advances in Business and Management Forecasting

Volume 5: Advances in Business and Management Forecasting

ADVANCES IN BUSINESS AND MANAGEMENT
FORECASTING VOLUME 6

ADVANCES IN BUSINESS AND MANAGEMENT FORECASTING

EDITED BY

KENNETH D. LAWRENCE
New Jersey Institute of Technology, Newark, USA

RONALD K. KLIMBERG
Saint Joseph's University, Philadelphia, USA

JAI

United Kingdom – North America – Japan
India – Malaysia – China

JAI Press is an imprint of Emerald Group Publishing Limited
Howard House, Wagon Lane, Bingley BD16 1WA, UK

First edition 2009

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN: 978-1-84855-548-8
ISSN: 1477-4070 (Series)

Awarded in recognition of Emerald's production department's adherence to quality systems and processes when preparing scholarly journals for print

INVESTOR IN PEOPLE

CONTENTS

This page intentionally left blank

LIST OF CONTRIBUTORS

Rebecca Abraham	Huizenga School of Business, Nova Southeastern University, Fort Lauderdale, FL, USA
Kallol Bagchi	Department of Information and Decision Sciences, University of Texas El Paso, El Paso, TX, USA
John E. Boylan	School of Business and Management, Buckingham Shire Chilton University College, Buckinghamshire, UK
Yu-Lin Chang	Department of Accounting and Information Technology, Ling Tung University, Taiwan
An-Sing Chen	College of Management, National Chung Cheng University, Ming-Hsiung, Chia-Yi, Taiwan
Huijing Chen	Salford Business School, University of Salford, Salford, UK
Shaw K. Chen	College of Business Administration, University of Rhode Island, RI, USA
Chung-Jen Fu	College of Management, National Yunlin University of Science and Technology, Yunlin, Taiwan
Charles Harrington	Huizenga School of Business, Nova Southeastern University, Fort Lauderdale, FL, USA

Christopher M. Keller Department of Marketing and Supply Chain Management, College of Business, East Carolina University, Greenville, NC, USA

Peeter Kirs Department of Information and Decision Sciences, University of Texas El Paso, El Paso, TX, USA

Ronald K. Klimberg DSS Department, Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA

John F. Kros Department of Marketing and Supply Chain Management, College of Business, East Carolina University, Greenville, NC, USA

N. K. Kwak Department of Decision Sciences and ITM, Saint Louis University, St. Louis, MO, USA

Tanya Lal Haub School of Business, Saint Joseph's University, Philadelphia, PA, USA

Kenneth D. Lawrence School of Management, New Jersey Institute of Technology, Newark, NJ, USA

Sheila M. Lawrence Management Science and Information Systems, Rutgers Business School, Rutgers University, Piscataway, NJ, USA

Chang Won Lee School of Business, Hanyang University, Seoul, Korea

Mark T. Leung Department of Management Science, College of Business, University of Texas at San Antonio, San Antonio, TX, USA

J. Gaylord May Department of Mathematics, Wake Forest University, Winston-Salem, NC, USA

Virginia M. Miori	DSS Department, Haub School of Business, St. Joseph's University, Philadelphia, PA, USA
Amitava Mitra	Office of the Dean and Department of Management, College of Business, Auburn University, Auburn, AL, USA
Daniel E. O'Leary	Levanthal School of Accounting, Marshall School of Business, University of Southern California, CA, USA
Dinesh R. Pai	Management Science and Information Systems, Rutgers Business School, Rutgers University, Newark, NJ, USA
Jayprakash G. Patankar	Department of Management, The University of Akron, Akron, OH, USA
Rolando Quintana	Department of Management Science, College of Business, University of Texas at San Antonio, San Antonio, TX, USA
Eddie Rhee	Department of Business Administration, Stonehill College, Easton, MA, USA
Gary J. Russell	Department of Marketing, Tippie College of Business, University of Iowa, Iowa City, IA, USA
Zaiyong Tang	Marketing and Decision Sciences Department, Salem State College, Salem, MA, USA
Joanne S. Utley	School of Business and Economics, North Carolina A&T State University, Greensboro, NC, USA
Frenck Waage	Department of Management Science and Information Systems, University of Massachusetts Boston, Boston, MA, USA

This page intentionally left blank

EDITORIAL BOARD

Editors-in-Chief

Senior Editors

This page intentionally left blank

PART I
FINANCIAL APPLICATIONS

This page intentionally left blank

COMPETITIVE SET FORECASTING IN THE HOTEL INDUSTRY WITH AN APPLICATION TO HOTEL REVENUE MANAGEMENT

John F. Kros and Christopher M. Keller

INTRODUCTION

Successful revenue management programs are found in industries where managers can accurately forecast customer demand. Airlines, rental car agencies, cruise lines, and hotels are all examples of industries that have been associated with revenue management. All of these industries have applied revenue management, whether it be complex overbooking models in the airline industry or simple price discrimination (i.e., having a tiered price system for those making reservations ahead of time versus walk-ups) for hotels.

The travel and hospitality industry and the hotel industry in particular has a history of employing revenue management to enhance profits. The ability to accurately set prices in response to forecasted demand is a central management performance tool. Individual hotel manager performance is generally base-lined on historical performance and a manager may be tasked with driving occupancy higher than the previous year's performance. In a stable market environment, such benchmarking is reasonable. However, if the market changes by the entrance and exit of competitive hotels then such

Advances in Business and Management Forecasting, Volume 6, 3–14

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006001

benchmarking may be unfair to the individual manager and a poor plan for management. This chapter develops two models that forecast demand and demonstrates for an existing data set how the entrance and exit of competitive market hotels in some cases does and in some cases does not change that forecasting demand. Understanding that the analysis of a dynamic market environment sometimes necessitates a forecast change is useful for improving the demand forecasts, which are critical to any revenue management system.

HOTEL REVENUE MANAGEMENT AND PERFORMANCE

The performance of hotel managers is measured along three principal components: revenue, cost, and quality. Hotel managers set availability restrictions and price for hotel rooms. Availability and price are the principal components of hotel revenue management. Of the three components of hotel management performance, revenue management is the single most controllably variable measure. Lee (1990) reported that accurately forecasting demand is cornerstone to any revenue management system and that a 10 percent improvement in forecast accuracy could result in an increase in revenue of between 1.5 and 3.0 percent.

Most hotel managers do not control fixed investment costs and generally speaking any controllable variable costs tend to be relatively standard rates. Cost management thus does not have large managerial flexibility, but rather is a limited control responsibility for overall hotel performance. Quality measures of performance include customer satisfaction and quality inspections. As with costs, measures of hotel quality performance are generally not widely variable, but rather are a limited control responsibility for overall hotel performance.

A number of researchers have studied hotel revenue management in general. Bitran and Mondschein (1995) spoke about room allocation, Weatherford (1995) proposed a heuristic for booking of customers, Baker and Collier (1999) compare five booking control policies. More specifically, forecasting in hotel revenue management has also been studied. Kimes (1999) studied the issue of hotel group forecasting accuracy and Weatherford and Kimes (2003) compare forecasting methods for hotel revenue management. Schwartz and Cohen (2004) investigate the subjective

estimates of forecast uncertainty by hotel revenue managers, whereas Weatherford, Kimes, and Scott (2001) speak more quantitatively to the concept of forecasting aggregation versus disaggregation in hotel revenue management. Weatherford et al. (2001) determine that disaggregated forecasts are much more accurate than in their aggregate form. Under their study, aggregation is done over average daily rate (ADR) and length-of-stay.

Rate class and length-of-stay are important variables that influence hotel performance and in turn improved forecasting of these variables should assist in optimizing revenue. Along these dimensions, ceteris paribus, hotel management is improved by increasing the performance measure. For example, higher rate classes generate higher revenue and longer lengths-of-stay generate greater revenue.

COMPARING REVENUE PERFORMANCE OF COMPETING HOTELS

The revenue performance of an individual hotel's revenue performance can be compared with its competitors' revenue performance using standard market data such as that provided by Smith Travel and Research (STAR). STAR reports are a widely used standard competitive market information service in the hotel industry. In a stable and nonchanging market, such comparisons can be used to assess individual performance of a hotel manager. But, in many markets, new hotels are opened, old hotels are closed, and existing hotels are rebranded (up or down). The dynamics of these market changes are generally reflected in the STAR reports. These dynamic market changes can seriously affect comparisons of an individual hotel and its competitive set.

The specific inclusion and exclusion of any hotel within the competitive set may vary over time and the entrance and exit of hotels in a competitive set is reflected in the STAR reports. This chapter considers an exploratory assessment of whether or not a competitive set has been well-defined for a specific hotel and in turn how a competitive set could be forecast for a specific hotel in a dynamic marketplace. The data consists of competitive performance data for five distinct periods, representing changes in hotel competitors. The performance measures of the dynamic market are considered for whether or not the underlying competitive market has

significantly changed and this is particular important for assessing individual hotel manager performance.

RESEARCH METHOD AND HOTEL DATA

This research examines hotel occupancy (OCC) and ADRs over a three-year period. The principal observation that is being sought is that ADR and OCC exhibit a consistent relationship over the various time periods. That is, the demand for hotel rooms has not changed during the time period; only the supply has changed by the entrance and or exit of hotels in the competitive set. There has been research in the area regarding model-based forecasting of hotel demand. Witt and Witt (1991b) present a literature review of published papers on forecasting tourism demand. They also compare the performance of time series and econometric models (Witt & Witt, 1991a). S-shaped models, based on Gompertz' work (Harrington, 1965), were used by Witt and Witt (1991a), and again by Riddington (1999) to model tourism demand. However, little other research has been completed incorporating such a model into the tourism area.

The data in the present chapter has five distinct periods over approximately three years with each period containing a different competitive set

(1) Period One: Data for an 8-month period, which includes 7 hotels with a total of 763 rooms. This variable is denoted P_1.
(2) Period Two: Data for an 11-month period, which includes 8 hotels with a total of 865 rooms, and includes the addition of a newly constructed brand name hotel. This variable is denoted P_2.
(3) Period Three: Data for a 5-month period, which includes 9 hotels with a total of 947 rooms, and includes the addition of a newly constructed brand name hotel. This variable is denoted P_3.
(4) Period Four: Data for a 1-month period, which includes 8 hotels with a total of 761 rooms, and includes the closing of an existing hotel. This variable is denoted P_4.
(5) Period Five: Data for a 10-month period, which includes 7 hotels with a total of 666 rooms, and includes the removal of an existing hotel from reporting within the competitive set. This variable is denoted P_5.

The periods as listed above are sequential. That is, Period Two follows Period One, and Period Five follows Period Four. It should be noted that each time period does not include each month, and that in general, each

month is included in only two or three of the relevant reporting periods. The nonsystematic and nonseasonal changes in the market limit the standard tools that may be applied to the data. For example, one method for considering the periodic effects above would be to construct a multiple linear regression model of demand that includes dummy variables for each of the periods. Although the dummy variable methodology for evaluating changes in the market is useful, the underlying linear model is difficult to apply reasonable in the data for this case of hotel revenue management. The fundamental issue with a linear approximation of demand is illustrated in Fig. 1. Descriptively, the "problem" is that any multiple linear regression model of demand is fundamentally ill-fitting.

This chapter uses a bases S-shaped model that begins with a slow start followed by rapid growth, which tails off as saturation approaches. As can be seen in Fig. 1, when ADR versus OCC is analyzed, the relationship definitely takes on an S-shape. Therefore, the basic Gompertz function was employed as the base function for modeling ADR versus OCC and is as follows:

$$y(t) = ae^{be^{ct}}$$

where $y(t) = \text{EstOCC}$, $a = 100$ (upper limit of % occupancy), $b < 0$ and is a curve fitting parameter, $c < 0$ and is a measure of growth rate, and $t = \text{ADR}$.

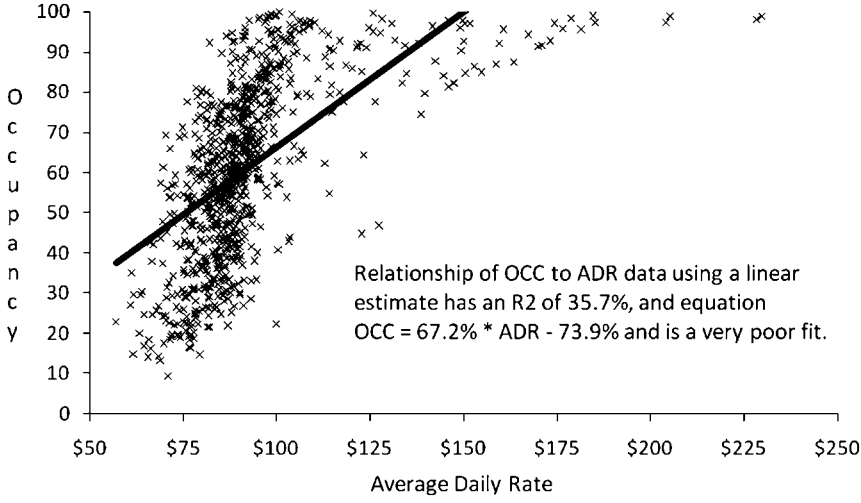

Fig. 1. Linear Estimate of Demand for the Data.

Since occupancy and price arise simultaneousy, the model does not state that occupancy is caused by price, but that as occupancy is expected to rise, prices rise. The occupancy model is then

$$\text{EstOCC}_i = 100 * e^{be^{ct}}$$

The model is shown in Fig. 2. As suggested previously, the possible structural effects of the varying periods can then be investigated by adding dummy variables representing each period to the model of occupancy based on the price-estimated occupancy as shown in Fig. 2.

The first period is the baseline period and is included in the intercept. The variable $P_2 = 1$, for all periods of time after the construction of the first new hotel, and is 0 in the first period only. This is a persistent variable, which continues throughout subsequent times. Similarly, $P_i = 1$, if the reporting period is greater than or equal to the present time, and otherwise equals 0, for each reporting period i.

The model for occupancy to be estimated is

$$\text{Occ}_i = \alpha_0 + \alpha_1 * \text{EstOCC}_i + \alpha_2 * P_2 + \alpha_3 * P_3 + \alpha_4 * P_4 + \alpha_5 * P_5$$

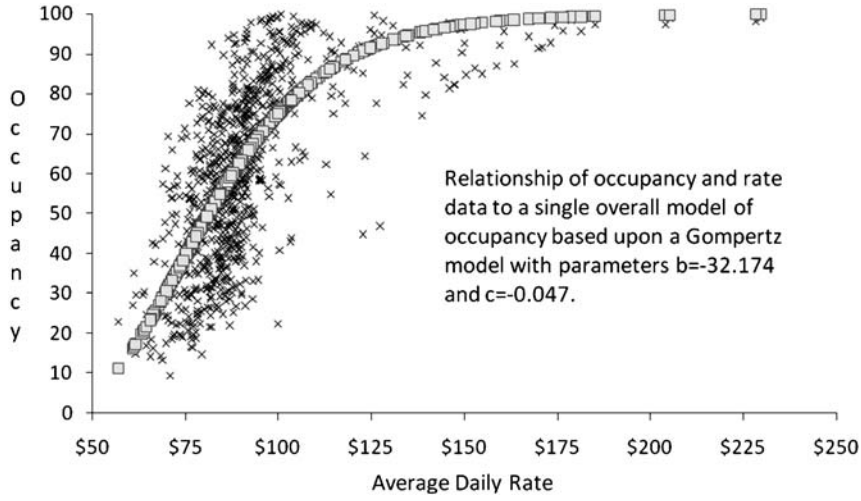

Fig. 2. Overall Occupancy Estimate using Gompertz Model.

The estimated parameters are

$$\begin{aligned}\text{Occ}_i = {} & 9.18 + 1.10 * \text{EstOCC}_i - 13.78 * P_2 - 10.96 * P_3 \\ & + 11.01 * P_4 - 3.84 * P_5\end{aligned}$$

The *t*-statistics for the coefficient estimates are, respectively, 4.38, 32.19, −7.83, −7.75, 3.78, and −1.37. All of the coefficient estimates are statistically significant at the 5% level with the exception of the fifth period, which has a *p*-value of 17%. Period 5 represented the removal of an existing hotel from the competitive data set and appears to not statistically significantly affect the overall estimation. That is, Period 5 represented a mere reporting change and not an actual supply change in the market.

Interestingly, each of the other three changes did significantly affect the estimate: two new constructions and the closing of an existing hotel. The overall results strongly suggest that the data as reported may in fact indicate underlying changes in the market reported. As a consequence, conclusions regarding hotel performance across the entire reporting period may not be directly justified and managerial performance structures that are based strictly on historical performance are not justified.

Since P_2 and P_3 represent increases in supply, then it is to be expected that the coefficient estimates are negative. Since P_4 represents a decrease in supply, then it is to be expected that the coefficient estimate is positive. The statistical insignificance of the coefficient for P_5 indicates that a mere reporting change in the data does not affect the estimate of occupancy, and that this variable may be removed from the estimated model.

A reduced model without this variable changes only slightly

$$\text{Occ}_i = 9.28 + 1.10 * \text{EstOCC}_i - 13.75 * P_2 - 10.95 * P_3 + 7.51 * P_4$$

The *t*-statistics for the coefficient estimates are, respectively, 4.43, 32.14, −7.81, −7.74, and 5.31, all of which are statistically significant beyond the 5% level. These results are shown below Fig. 3.

The basic model shown above does demonstrate that competitive market changes appear to significantly impact estimates of occupancy. As a result measuring individual hotel manager performance on a simple historical basis is unfair to the manager and since it is inconsiderate of the changing market will represent a poor performance plan for the hotel management as it is fundamentally at odds with reality. Having noted this change, however, it may be also necessary to consider the interaction effects between the months or seasons that are included within each period. Although the simple model does indicate that changes in supply cause changes in

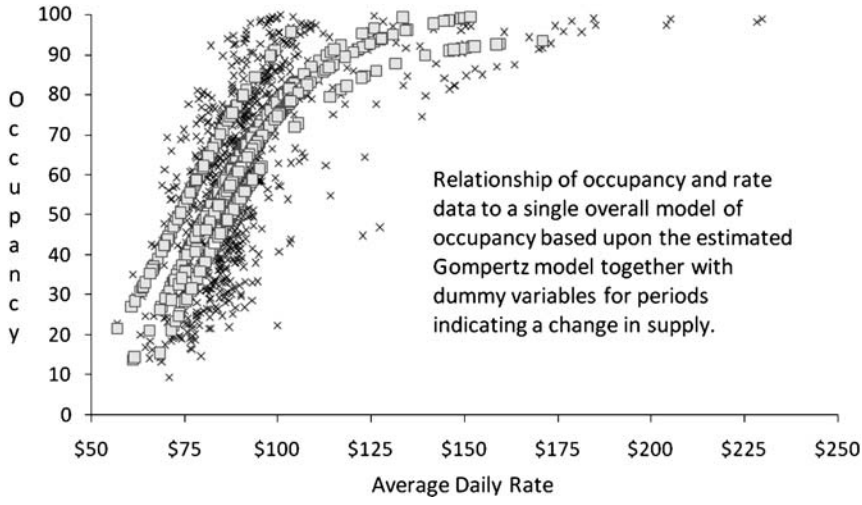

Fig. 3. Estimated Relationship with Supply Period Dummy Variables.

occupancy rates, because the periods include widely varying periods of time, it may be the case that the significance of the period dummy variables is a result not of the supply changes, but perhaps is a result that masks the underlying demand changes associated with the included months of each period. To address this complexity of interaction effects between a changing market and changing demand during the changing market periods, a second model of occupancy is constructed that specifically considers occupancy changes by month.

Since there are seasonal or monthly effects in demand estimation, the result above might be an implicit consequence that the periods mask underlying monthly or seasonal demand changes. The underlying estimate of demand is estimated separately for each month and the results of the prediction parameters are shown in Table 1.

It is clear from looking at Table 1 that the parameter estimates may vary greatly over the individual months. Once these factors are considered, then a composite model may be constructed that considers whether or not the effects are different. This model consists of 12 models, one for each month j, but with single parameters across the various periods comprising multiple months

$$\begin{aligned}\text{Monthly}_j\text{Occ}_i = {} & \alpha_0 + \alpha_1 * \text{Monthly}_j\text{EstOCC}_i \\ & + \alpha_2 * P_2 + \alpha_3 * P_3 + \alpha_4 * P_4 + \alpha_5 * P_5\end{aligned}$$

Table 1. Gompertz Parameter Estimates for Individual Months.

	b	*c*
January	−34.839	−0.046
February	−749.134	−0.084
March	−591.507	−0.082
April	−45.909	−0.052
May	−4.744	−0.026
June	−1,803.937	−0.095
July	−29.490	−0.050
August	−5.962	−0.030
September	−7.495	−0.028
October	−8.798	−0.032
November	−174.096	−0.063
December	−6.716	−0.025

The estimated parameters are

$$\text{Monthly}_j\text{Occ}_i = 6.22 + 1.03 * \text{Monthly}_j\text{EstOCC}_i - 8.14 * P_2 - 2.00 * P_3 + 0.63 * P_4 + 0.71 * P_5$$

In this case, the reporting variable in P_5 is once again statistically insignificant. More importantly however is that more than one of the variables is statistically insignificant at the 5% level. The respective *t*-statistics are 3.11, 35.49, −5.01, −1.48, 0.23, and 0.27. Utilizing a backwards elimination regression with the least significant removed at each step, the following model attains: eliminate P_4; eliminate P_5; and eliminate P_3: $\text{Monthly}_j\text{Occ}_i = 5.98 + 1.03 * \text{Monthly}_j\text{EstOCC}_i - 8.89 * P_2$. The respective *t*-statistics are 3.08, 37.90, and −5.87, all significant beyond the 5% level.

The final composite model in addition to the one remaining dummy variable is shown in Fig. 4. It illustrates the greater flexibility of the model estimates.

The data conclusions from this model are very interesting. Once monthly variation is included in the model, then only one of the market reporting changes affects estimated occupancy. Period 5, which represented a mere reporting change in the hotels included within the competitive set, had no statistical effect on estimated occupancy. This result is perhaps expected. Period 4, which represented the closing of an existing hotel, had a substantial decrease in supply. As a first analysis, one would expect that a decrease in supply would substantially increase the overall occupancy of the other hotels. However, the model data shows that the closing of the hotel does not substantially affect the overall occupancy of the other hotels. The secondary

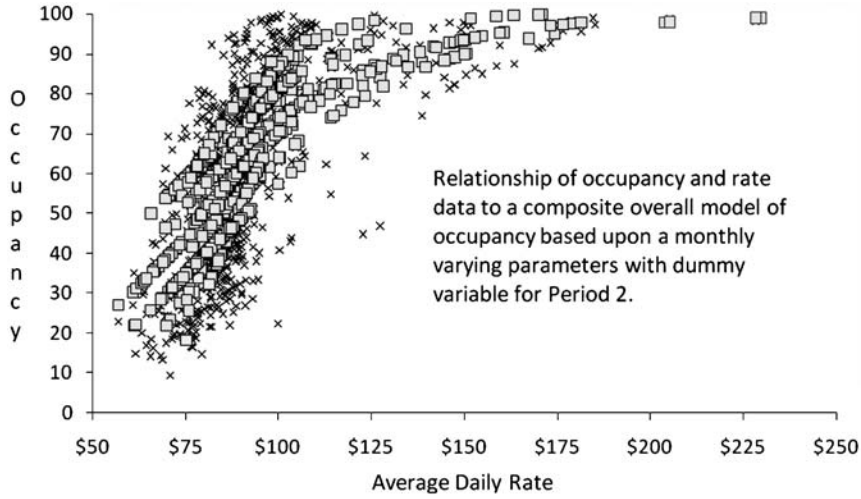

Fig. 4. Monthly Model Variation with Single Dummy Estimate for Period 2.

dynamic explanation is that the hotel closed because it was not effective at garnering occupancy and so its closure did not have a substantial ripple effect on the other hotels. Period 3, which represented the construction of a new hotel, would imply an expectation that this increase in supply would substantially decrease overall hotel occupancy percentage. The model data shows that this new construction does not substantially affect the overall occupancy of the other hotels and that this new hotel may have been an overbuild subsequently facing financial difficulty. Finally, the data does show that the construction of this new hotel did impact overall hotel occupancy. This hotel was a successful market entrant that changed the allocation of expected occupancy to all other rooms. The effect of this hotel was substantial, creating nearly a 9% decrease in overall hotel occupancy percentage because of the increase in supply. This change and the absence of changes are very important for evaluating individual hotel manager performance and for setting management performance goals in the market, especially if managerial performance management is generally structured on past comparative performance.

MANAGERIAL IMPLICATIONS AND CONCLUSIONS

Revenue management typically has wide flexibility and significant responsibility for overall hotel performance. Revenue management is composed of the

product of the occupancy and the price. If the demand relationship of customers does not change over time and this is reasonable because hotel managers act as a set of revenue-maximizers. This means that regardless of the supply, the price–demand relationship will remain unchanged. Because of this, then a principal dimension for managerial performance enhancement is estimating the occupancy in relation to price and individual performance targets for hotel managers may be base-lined on historical performance. That is, a hotel manager may be tasked with driving occupancy above the previous year's performance. In a stable market environment, such benchmarking is reasonable. However, if the market changes by the entrance or exit of competitive hotels, then such benchmarking may be unfair to the manager and a poor plan for the management. This chapter develops a method for analyzing then the entrance and exit of hotels changes the competitive market environment and the attendant effects on occupancy.

One of the things that remains unknown for future research is can a market entrant or exit be accurately predicted of whether or not it will in the future, rather than in the past, affect performance. That is, this chapter evaluated ex-post the market changes which is useful for evaluating performance. In terms of future investment in new properties, what would be especially valuable is a determination of whether or not the addition of supply would directly impact occupancy estimates.

Although historic comparisons are useful to assessing changing performance, in order to assess competitive performance it is necessary to compare hotel revenue performance to the hotel's competitors' revenue performance. In a broader sense, hotel managers are also interested in their relative position and performance to competitors in their market. The decomposition of the relevant data into comparative changes for individual hotels within the present data set may also be a valuable avenue of future research. In any case, this chapter has developed and illustrated a method for considering occupancy effects on existing hotels with attendant changes in the market supply illustrating both the existence of an increase in supply that cause a decrease in overall occupancy, an increase in supply that does not cause a decrease in occupancy, and a decrease in supply that does not result in an increase in occupancy.

REFERENCES

Baker, T. K., & Collier, D. A. (1999). A comparative revenue analysis of hotel yield management heuristics. *Decision Sciences*, *30*(1), 239–263.

Bitran, G. R., & Mondschein, S. (1995). An application of yield management to the hotel industry considering multiple days stays. *Operations Research*, *43*, 427–443.

Harrington, E. C. (1965). The desirability function. *Industrial Quality Control*, *21*, 494–498.

Kimes, S. (1999). Group forecasting accuracy for hotels. *Journal of the Operational Research Society*, *50*(11), 1104–1110.

Lee, A. O. (1990). *Airline reservations forecasting: Probabilistic and statistical model of the booking process*. Ph.D. Thesis, Massachusetts Institute of Technology.

Riddington, G. L. (1999). Forecasting ski demand: Comparing learning curve and varying parameter coefficient approaches. *Journal of Forecasting*, *18*, 205–214.

Schwartz, Z., & Cohen, E. (2004). Hotel revenue-management forecasting: Evidence of expert-judgment bias. *Cornell Hotel and Restaurant Administration Quarterly*, *45*(1), 85–98.

Weatherford, L. R. (1995). Length-of-stay heuristics: Do they really make a difference? *Cornell Hotel and Restaurant Administration Quarterly*, *36*(6), 70–79.

Weatherford, L. R., & Kimes, S. E. (2003). A comparison of forecasting methods for hotel revenue management. *International Journal of Forecasting*, *19*(3), 401–415.

Weatherford, L. R., Kimes, S. E., & Scott, D. (2001). Forecasting for hotel revenue management: Testing aggregation against disaggregation. *Cornell Hotel and Restaurant Administration Quarterly*, *42*(6), 156–166.

Witt, S., & Witt, C. (1991a). Tourism forecasting: Error magnitude, direction of change error and trend change error. *Journal of Travel Research*, *30*, 26–33.

Witt, S., & Witt, C. (1991b). Forecasting tourism demand: A review of empirical research. *International Journal of Forecasting*, *11*, 447–475.

PREDICTING HIGH-TECH STOCK RETURNS WITH FINANCIAL PERFORMANCE MEASURES: EVIDENCE FROM TAIWAN

Shaw K. Chen, Chung-Jen Fu and Yu-Lin Chang

ABSTRACT

A one-year-ahead price change forecasting model is proposed based on the fundamental analysis to examine the relationship between equity market value and financial performance measures. By including book value and six financial statement items in the valuation model, current firm value can be determined and the estimation error can predict the direction and magnitude of future returns of a given portfolio. The six financial performance measures represent both cash flows – cash flows from operations (CFO), cash flows from investing (CFI), and cash flows from financing (CFF) – as well as net income – R&D expenditures (R&D), operating income (OI), and adjusted nonoperating income (ANOI). This study uses a 10-year sample of the Taiwan information electronic industry (1995–2004 with 2,465 firm-year observations). We find hedge portfolios (consisting of a long position in the most underpriced portfolio and an offsetting short position in the most overpriced portfolio) provide an average annual return of 43%, more than three times the average annual stock return of 12.6%. The result shows the estimation error can be a

Advances in Business and Management Forecasting, Volume 6, 15–35

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006002

good stock return predictor; however, the return of hedge portfolios generally decreases as the market matures.

1. INTRODUCTION

Fundamental analysis research is aimed at determining the value of firm securities by carefully examining critical value drivers (Lev & Thiagarajan, 1993). The importance of analyzing the components of financial statements in assessing firm value and future stock returns is widely highlighted in financial statement analyses. Beneish, Lee, and Tarpley (2001) assert securities with a tendency to yield unpredictable returns are particularly attractive to professional fund managers and investors and suggest fundamental analysis based on financial performance measures is more helpful, in terms of correlation with future returns. Therefore, providing evidence on the association between financial performance measures and contemporaneous stock prices or future price changes is an important issue for academics and practice.

The evidence from researches shows stock prices do not completely reflect all publicly available information (Ou & Penman, 1989a; Fama, 1991; Sloan, 1996; Frankel & Lee, 1998). Followed by Ball and Brown (1968) and Beaver (1968), many studies examine the association with stock returns to compare alternative accounting performance measures, such as earnings, accruals, operating cash flows, and so on. However, numerous prior valuation researches suggest the market might be informationally inefficient and stock prices might take years before they entirely reflect available information. This leads to significant abnormal returns spread over several years by implementing fundamental analysis trading strategies (Kothari, 2001).

The major goal of fundamental analysis is to assess firm value from financial statements (Ou & Penman, 1989a). Using the components of financial statements to help estimate a firm's intrinsic value, the difference between the intrinsic value and the stock price can be examined whether it successfully identifies those mispriced securities. Related research such as Ou and Penman (1989a, 1989b), Lev and Thiagarajan (1993), Abarbanell and Bushee (1997, 1998), and Chen and Zhang (2007) predict future earnings and stock returns using financial measures within the income statement and balance sheet. However, a number of studies present evidence investors do not correctly use available information in predicting future earnings performance (Bernard & Thomas, 1989, 1990; Hand, 1990; Maines &

Hand, 1996). Furthermore, the research in return predictability also provides strong evidence challenging market efficiency (Ou & Penman, 1989a, 1989b; Holthausen & Larcker, 1992; Sloan, 1996; Abarbanell & Bushee, 1997, 1998; Frankel & Lee, 1998).

In practice, the purpose of fundamental analysis research is to identify mispriced securities for investment decisions. Several empirical studies document intrinsic values estimated using the residual income model to help predict future returns (Lee, Myers, & Swaminathan, 1999). Kothari (2001) indicates the research on indicators of market mispricing produce large magnitudes of abnormal returns and further suggests a good model of intrinsic value should predictably generate abnormal returns. This raises the issue of how to precisely predict stock returns with financial performance measures based on the residual income model. Testing what kind of financial performance measures should be embedded in the valuation model is also important to investors.

Accruals and cash flows are the two financial measures most commonly examined. Prior related studies compared stock returns' association with earnings, accruals, and cash flows, such as Rayburn (1986), Bowen, Burgstahler, and Daley (1986, 1987), and Bernard and Stober (1989). This study extends the research for value-relevant fundamentals to investigate how the major components of financial statements enter the decisions of market participants, and we highlight the research design correlating the unanticipated component with stock returns.

To perform fundamental analysis to examine the relationship between equity market value and key financial performance measures, we use book value as the basic firm value measure and further decompose cash flows into cash flows from operations (CFO), cash flows from investing (CFI), and cash flows from financing (CFF), and net income into R&D expenditures (R&D), operating income (OI), and adjusted nonoperating income (ANOI) as additional performance measures. Applying these seven financial performance measures (CFO, CFI, CFF, R&D, OI, ANOI and book value) with the stock price, this study constructs fundamental valuation models, and the estimated errors can be used to predict future stock returns. Then, this study operates hedge portfolios of major high-tech companies in Taiwan based on ranked estimation errors involving a long position in the most underpriced stocks and an offsetting short position in the most overpriced stocks. The empirical results show the highest average one-year holding-period return of the hedge portfolios is about 43%, much higher than the risk-free rate and average stock return (12.6%) in the same period. Our finding suggests the estimation error of valuation model embedded in

these seven financial performance measures can be a good stock return predictor.

This study adds to the growing body of evidence indicating stock prices reflect investors' expectations about fundamental valuation attributes such as earnings and cash flows. In particular, it further contributes in two respects. First, we apply and confirm a model relying on characteristics of the underlying accounting process that are documented in texts on financial statement analysis; second, our empirical findings suggest major components of earnings and cash flows will lead to better predictions of relative future cross-sectional stock returns.

The rest of this chapter is organized as follows. Section 2 discusses the literature review and develops the hypotheses. Section 3 discusses the estimation procedures used to implement the valuation model and contains the sample selection procedure, definition and measurement of the variables, respectively. Section 4 discusses the empirical results and analysis. Final section concludes with a summary of our findings and their implications.

2. LITERATURE REVIEW AND HYPOTHESES

Shareholders, investors, and lenders have an interest in evaluating the firm for decision making. Firm's current performance will be summarized in its financial statements, and then the market assesses this information to evaluate the firm. This is consistent with the conceptual framework of Financial Accounting Standard Board (FASB) that financial statements should help investors and creditors in "assessing the amounts, timing, and uncertainty of future cash flows" (Kothari, 2001). A growing body of related research shows stock prices do not fully reflect all publicly available information. It implies that investor might irrationally decide the stock price. Many studies further evaluate the ability of the models to explain stock prices. Penman and Sougiannis (1998) examine variations of the model by means of ex-post realizations of earnings to proxy for ex ante expectations. Dechow, Hutton, and Sloan (1999) investigate the models under alternative specifications and examine the predictive power of the models for cross-sectional stock returns in the US. Dechow (1994) further indicates prior research emphasizing on unexpected components of the financial performance measures is misplaced, and suggests researchers should search for the best alternative measure in the valuation model to evaluate firm value.

The evidence in prior research indicates earnings and cash flows are value-relevant. The relation between stock prices and earnings has been widely researched. In view of components of earnings, Robinson (1998) indicates OI is a superior measure for firms to reflect the firm's ability to sell its products and evaluate firm's operation performance. Furthermore, if the non-productive and idle assets are disposed and managers make decisions in investor's best interest, Fairfield, Sweeney, and Yohn (1996) suggest ANOI has incremental predictive content of future profitability. Thus, both OI and ANOI have a positive association with firm value. R&D is a particularly critical element in the production function of firms in the information electronic industry. Evidence supports that R&D should be included as an important independent variable in the valuation model for research-intensive firms. Hand (2003, 2005) finds equity values are positively related to R&D and supports the proposition that R&D is beneficial to the future development of the firm. Chan, Lakonishok, and Sougiannis (1999) further suggest R&D intensity is associated with return volatility. In sum, as the components of earnings increases, such as OI, ANOI, and R&D, the value of the firm will increase.

For cash flows, Black (1998), Jorion and Talmor (2001), and Hand (2005) suggest that the CFI are value relevant. Firms wanting to increase their long-term competitive advantage will spend more on investment activities to engage in expansion and profit-generating opportunities. For this reason, firms will execute financial activities to raise sufficient funds excluding from operations. Jorion and Talmor (2001) suggest the ability to generate funds from the outside affects the opportunity for firms' continuous operations and improvement. When the CFF is higher, it has a direct benefit on the value of the firm. Furthermore, CFO, as a measure of performance is less subject to distortion than the net income figure. Klein and Marquardt (2006) views CFO as the measure of the firm's real performance.

In sum, the components of earnings or cash flows have different implications for assessing firm value. The evidence of prior studies support a positive contemporaneous association of earnings and cash flows with stock returns (or stock prices), which is generally attributed to earnings and cash flows to summarize value-relevant information.

The well-known Ohlson's (1995) residual income model decomposed the equity evaluation into book value, earning capitalization value, and present value of "other information." Ohlson (1995) states the importance of "other information" and further points out other information is difficult to measure and operate, on the assumption investors on average could use the financial statements to capture the main value of other information in an

unbiased manner. By identifying the role of information from the components of earnings and cash flows in the forecasting of future stock returns and firm value, this study provides an expected setting in which to support and extend prior research.

On the other hand, numerous recent studies conclude the capital market is inefficient with respect to some areas (Fama, 1991; Beaver, 2002).[1] In addition, Grossman and Stiglitz (1980) show the impossibility of information efficient markets when information is costly, and there is an equilibrium degree of disequilibrium, price reflects the information of informed individuals but only partially.

We expect the market perception of the relationship between the firm value and its components of financial statements on average are unbiased, but it will not always be completely reflected in stock prices in time for some individual securities. Thus, we can evaluate the relationship between various financial measures and firm value by regressing key components of financial statements to the stock prices. Therefore, we infer financial performance measure stated earlier such as the R&D, OI, ANOI, CFF, CFI, and CFO is increasing; it has a direct benefit on the value of the firm. By connecting stock price with book value and these six financial measures in the valuation model, we propose the estimation error of the fundamental value equation may serve an indicator of the direction and magnitude of future stock return and further examine its usefulness in predicting future stock returns in the Taiwan capital market.

This study expects incorporating more comprehensive and more precise (by focusing on the same industry) estimators should improve the predictability in estimating those firm's intrinsic values. So we can use the estimated coefficients of the variables in the regression to estimate each firm's "intrinsic values," and compare it with the actual stock price to identify overpriced and underpriced stocks to investigate the issue related to the predictability of future stock returns. From the evidence of prior research, at times stock price deviate from their fundamental values and over time will slowly gravitate toward the fundamental values. If the stock prices are gravitating to their fundamental values, then the trading strategy of a hedge portfolio formed by being on the long side of the underpriced stocks and on the short side of the overpriced stocks will produce "abnormal return."

In sum, we expect the trading strategy taking a long position in firms with negative estimation error and a short position in firms with positive estimation error should generate abnormal stock returns. In additions, when we operate the trading strategy with relatively higher absolute estimation errors, the more abnormal stock returns will be generated. Thus, we infer

the more the estimation errors in the current year, the less the subsequent stock returns.

Hence we state the hypothesis as follows:

H_1. There is a negative association between current year estimation errors and subsequent stock returns.

3. RESEARCH DESIGN AND METHODOLOGY

3.1. Empirical Model and Variables Measurements

To examine the research hypothesis raised above, we first select variables that reflect levels of firm values. Following the residual income valuation model of Ohlson (1995) and Feltham and Ohlson (1995), this study further modifies the prediction model employed by Frankel and Lee (1998), Dechow et al. (1999), and Lee et al. (1999)[2] to reexamine the relation between seven financial measures and firm value in the Taiwan capital market.

The valuation model to examine the relationship between equity market value and the various financial performance measures, decomposing cash flows into CFO, CFI, and CFF, and net income into R&D, OI, and ANOI, is presented as follows (for the definition and measurement of variables see Table 1):

$$\begin{aligned} MV_i = {} & b_0 + b_1 BVC_i + b_2 CFO_i + b_3 CFI_i + b_4 CFF_i + b_5 R\&D_i \\ & + b_6 OI_i + b_7 ANOI_i + \varepsilon_i \end{aligned}$$

Given the discussion above, these variables should be positively associated with expected stock prices,[3] except for CFI, which may be positively or negatively associated with expected stock prices, depending on the situation.

After we use these seven variables to regress the stock price, we can estimate each firm's fundamental values. Then, we use the error term (or estimation error) of the estimation equation to predict the next period stock return and expect there to be a negative relationship between the estimation error and subsequent stock return (or stock price change) and further construct the trading strategy to invest in those mispriced stocks as ranked by their estimated errors. The estimation procedure is described in more detail in Section 3.2.

Table 1. Definition and Measurement of Variables.

Variables	Measurement
Dependent variables	
Market value of equity (MV_{it})	The market value of equity of firm i at time t/AV_{t-1}
Independent variables	
Book value of net assets except for cash (BVC_{it})	The book value of equity less the change in the cash account of firm i at time t/AV_{t-1}
R&D expense ($R\&D_{it}$)	R&D expense of the firm i at time t/AV_{t-1}
Operating income (OI_{it})	(Gross profit−operating expenses) of firms i at time t/AV_{t-1}
Adjusted nonoperating income ($ANOI_{it}$)	$(NI_{it}-OI_{it}-R\&D_{it})/AV_{t-1}$
Cash flows from operations (CFO_{it})	Cash flows from operations activities of firm i at time t/AV_{t-1}
Cash flows from investing (CFI_{it})	Cash flows from investing activities of firm i at time t/AV_{t-1}
Cash flows from financing (CFF_{it})	Cash flows from financing activities of firm i at time t/AV_{t-1}
Other variables	
Operating income before discontinued and extraordinary items (NI_{it})	The operating income before discontinued and extraordinary items of firms i at time t/AV_{t-1}

Note: To allow for a cross-sectional aggregation and mitigate the impact of cross-sectional difference in firm size, we deflate all of the variables for each year by the book value of assets at the end of year $t-1(AV_{t-1})$.
Source: Fu and Chang (2006, p.17).

3.2. The Estimation Procedures

The purpose of this study is to examine the possibility of temporary stock mispricing that can be systematically predicted by our particular valuation models. Thus, we consider whether the estimation errors derived from the seven financial performance measures implied in the valuation model are able to predict future stock returns. The estimation procedures taken in this study are implemented as follows.

First, we assume the seven financial measures in the estimation model stated earlier provide important information for determining current firm's value and can help predict future stock return, with error terms providing the direction and magnitude of price changes. Thus, by regressing stock price to the seven financial measures (BVC, CFO, CFI, CFF, R&D, OI, and ANOI), we use the estimated coefficients of these variables in the regression and estimate each firm's "intrinsic values."

Second, compare each firm's "intrinsic values" with actual stock price to compute the error terms of valuation model and identify overpriced stocks (with positive error terms) and underpriced stocks (with negative error terms).

Third, using the relative value of error terms as the return predictor, we develop a trading strategy, providing insight into the deviations from the rational fundamental value expectations and actual stock prices.[4] The absolute value of expected stock returns (error terms) are ranked from low to high and then assigned to five hedge portfolios[5] (or three hedge portfolios) based on equal numbers. Therefore, quintile 1 portfolio is formed with the lowest ranked error terms; in contrast, quintile 5 portfolio is formed with the highest ranked error terms.

Finally, the study operates each hedge portfolio by going long on the underpriced stocks and shorting the overpriced ones, and then calculates the average expected future annual portfolio return for each quintile portfolio for the period of 1995–2004. So, for example, we expect the lowest quintile (Quintile 1) has the highest stock return average; in contrast, the highest quintile (Quintile 5) will have the lowest stock return average.[6]

In sum, by combining these variables in a prediction model, we develop an estimate of the error terms in yearly subsequent stock return forecasts, and show this estimate has predictive power for cross-sectional returns. Then, the investors can gain the excess return from these hedge portfolios, and assess the relationship between stock price and these financial perfo-rmance measures and use the estimation error as a benchmark to forecast subsequent stock prices changes or return.

3.3. Sample Selection

The sample companies are composed of publicly listed firms on the Taiwan Stock Exchange (TSE) and Gre Tai Securities Market (the GTSM) companies. The companies' financial data and the equity market value data are obtained from the financial data of company profile of the Taiwan Economic Journal (TEJ) Data Bank. The criteria for sample selection are as follows:

(1) Sample firms are limited to information electronics industries.
(2) For each year, companies without sufficient stock price or financial data are excluded from this study.
(3) Companies subject to full-delivery settlements and the de-listed companies are excluded.

Eliminating firms due to lack of sufficient data gives a sample size of 2,465 firm-years observations for the 10-year period from 1995 to 2004.

We focus on the information electronic industry in Taiwan for three reasons (1) The information electronic industry in Taiwan is the most important and competitive industry; (2) Beneish et al. (2001) suggest professional analysts typically tend to focus on firms with the same industry; and (3) We can mitigate some problems of cross-sectional studies (Ittner, Larcker, & Randall, 2003). Furthermore, by incorporating more complete value drivers and more precise estimates of their coefficients in the prediction model in the same industry, we could improve our ability to explain contemporaneous stock prices and therefore predict future stock returns.

4. EMPIRICAL RESULTS AND ANALYSIS

4.1. Descriptive Statistics

In Table 2, we present the descriptive statistics for the full sample and annual samples. As we show in panel A of Table 2, the mean (median) of MV are 18,680 (3,080) million New Taiwan dollars. The mean (median) of CFO, CFI, and CFF are 1,281 (138), −1,531 (−209), and 452 (49) million New Taiwan dollars, respectively. On the other hand, the mean (median) of R&D, OI, and ANOI are 269 (56), 785 (147), and –226 (–73) million New Taiwan dollars, respectively. The mean (median) of Return is 16.95% (−0.74%). Panel B of Table 2 shows the mean (or median) of these variables for each year are quite diverse; implying there are divergent characteristics among those firms.

Table 3 reports the Spearman and Pearson correlations among selected variables. Except for CFI, the other six financial performance measures are positively related to MV. As expected, in general, observed relations among variables are consistent with our expectations.

4.2. Empirical Results and Analysis

Our research purpose is to predict both the direction and the magnitude of deviations in the expectations of stock prices and to examine whether market value weighted average (or simple average) error terms in the

Table 2. Descriptive Statistics of the Sample.

Variable	MV	BVC	CFO	CFI	CFF	R&D	OI	ANOI	Return
Panel A: Descriptive statistics of the full sample (1995–2004)									
Mean	18,680	7,541	1,281	−1,531	452	269	785	−226	16.95%
25%	1,311	875	−4	−717	−88	19	32	−197	−35.26%
Median	3,080	1,722	138	−209	49	56	147	−73	−0.74%
75%	8,475	4,048	492	−45	446	166	414	−10	41.90%
Standard Deviation	76,379	26,038	7,152	6,598	3,668	885	4,180	1,462	92.46%
Observations	2,465	2,465	2,465	2,465	2,465	2,465	2,465	2,465	2,465
Panel B: Descriptive statistics of annual samples									
1995									
Mean	14,074	4,932	986	−1,655	1,220	201	1,196	−8	120%
Median	5,282	2,083	32	−318	404	80	196	−67	85%
Standard Deviation	25,410	7,791	3,474	3,570	2,365	352	2,924	627	99%
Observations	50	50	50	50	50	50	50	50	50
1996									
Mean	14,746	5,302	1,033	−1,766	973	210	816	−18	28%
Median	6,902	1,892	222	−285	136	66	194	−52	13%
Standard Deviation	25,745	10,141	3,056	4,730	2,468	377	2,300	921	56%
Observations	75	75	75	75	75	75	75	75	75
1997									
Mean	27,012	6,564	720	−1,989	1,711	235	614	225	−19%
Median	8,305	2,360	147	−608	570	80	196	11	−26%
Standard Deviation	59,200	12,012	2,705	4,691	2,833	414	1,802	914	34%
Observations	92	92	92	92	92	92	92	92	92
1998									
Mean	22,027	6,305	774	−1,880	1,297	225	452	−225	50%
Median	6,145	2,051	155	−581	566	38	191	−49	67%
Standard Deviation	50,626	10,630	2,532	4,472	2,258	401	1,781	854	239%
Observations	133	133	133	133	133	133	133	133	133
1999									
Mean	37,784	7,029	788	−1,513	875	231	712	−42	−38%
Median	6,561	2,045	141	−310	179	55	134	−59	−42%
Standard Deviation	121,220	15,612	3,791	4,574	2,566	492	2,504	908	26%
Observations	178	178	178	178	178	178	178	178	178
2000									
Mean	21,421	9,478	1,741	−2,478	1,198	289	1,406	−103	16%
Median	2,612	2,062	120	−323	3,873	54	181	−51	5%
Standard Deviation	97,927	32,804	8,824	8,462	257	837	6,159	1,411	49%
Observations	241	241	241	241	241	241	241	241	241

Table 2. (*Continued*)

Variable	MV	BVC	CFO	CFI	CFF	R&D	OI	ANOI	Return
2001									
Mean	23,841	8,575	1,469	−1,731	288	301	440	−408	−45%
Median	3,618	1,787	199	−224	−4	52	101	−66	−47%
Standard Deviation	102,019	30,621	6,758	6,540	3,446	1,020	3,632	1,721	21%
Observations	308	308	308	308	308	308	308	308	308
2002									
Mean	12,635	7,367	1,173	−1,255	219	268	611	−362	62%
Median	2,430	1,577	133	−184	5	53	121	−84	41%
Standard Deviation	52,846	26,349	6,976	5,583	3,064	879	3,579	1,879	76%
Observations	408	408	408	408	408	408	408	408	408
2003									
Mean	15,129	6,927	1,234	−998	72	265	689	−199	−23%
Median	2,594	1,525	113	−128	3	55	145	−70	−28%
Standard Deviation	70,759	25,204	7,370	4,517	3,899	839	3,920	1,341	45%
Observations	490	490	490	490	490	490	490	490	490
2004									
Mean	14,650	8,011	1,566	−1,480	−15	293	1,041	−295	43%
Median	2,011	1,527	139	−103	−12	61	161	−111	26%
Standard Deviation	69,614	35,201	9,138	9,482	4,632	892	5,403	1,518	71%
Observations	490	490	490	490	490	490	490	490	490

Notes: The variables are presented in millions New Taiwan dollars and are not deflated by the book value of assets at the end of year $t-1$(AV_{t-1}). Variables MV, BVC, CFO, CFI, CFF, R&D, OI, and ANOI are defined in Table 1. Return is the annual stock return at the end of April.

Table 3. Correlation Coefficients among Variables.

Variable	MV	BVC	CFO	CFI	CFF	R&D	OI	ANOI
MV	1	0.666***	0.362***	−0.453***	0.206***	0.329***	0.701***	0.145***
BVC	0.534***	1	0.243***	−0.499***	0.153***	0.258***	0.500***	0.185***
CFO	0.339*	0.165***	1	−0.309***	−0.300***	0.126***	0.467***	0.001
CFI	−0.391***	−0.516***	−0.237***	1	−0.591***	−0.115***	−0.389***	−0.093***
CFF	0.280***	0.336***	−0.226***	−0.754***	1	0.021***	0.100***	0.103***
R&D	0.361***	0.240***	0.174***	−0.158***	0.107***	1	0.228***	−0.432***
OI	0.640***	0.454***	0.483***	−0.357***	0.188***	0.295***	1	−0.072***
ANOI	0.074***	0.208***	0.020***	−0.089***	0.082***	−0.326***	0.015	1

Notes: Right-up, Spearman Correlation; left-down, Pearson Correlation; Variables MV, BVC, CFO, CFI, CFF, R&D, OI, and ANOI are defined in Table 1.
***Significant at 1% level; *significant at 10% level (two-tailed).

valuation model have predictability in future returns of the five (or three) hedge portfolios.

Panel A of Table 4 shows the results of the quintile hedge portfolios' return-predictability performance. We find the average excess return for the

Table 4. Estimation Results of Hedge Portfolios with Respect to Future Annual Stock Returns Sample Consists of 2,465 Firm-Years between 1995 and 2004.

Portfolios	MV_t	MVWA (Apr) (%)	Portfolios	MV_t	MVWA (Apr) (%)
Panel A: 5 hedge portfolios			Panel B: 3 hedge portfolios		
Quintile 1 (lowest error terms)			*Tercile 1 (lowest error terms)*		
Mean	15,639.72	42.25	Mean	13,216.84	37.82
Standard Deviation	38,126.91		Standard Deviation	34,925.24	
Median	5,851.20		Median	4,816.55	
Observations	493		Observations	822	
Quintile 2			*Tercile 2*		
Mean	8,523.50	30.63	Mean	11,894.26	17.88
Standard Deviation	16,192.59		Standard Deviation	28,110.95	
Median	3,830.10		Median	4,053.50	
Observations	493		Observations	822	
Quintile 3			*Tercile 3 (highest error terms)*		
Mean	12,967.22	18.55	Mean	35,634.03	4.01
Standard Deviation	25,321.97		Standard Deviation	95,572.92	
Median	5,138.70		Median	8,384.00	
Observations	493		Observations	821	
Quintile 4					
Mean	19,491.02	15.99			
Standard Deviation	50,905.82				
Median	5,208.90				
Observations	493				
Quintile 5 (highest error terms)					
Mean	44,572.83	−0.69			
Standard Deviation	113,556.98				
Median	7,586.30				
Observations	493				
Full sample			*Full sample*		
Mean	20,267.32	12.6	Mean	20,267.32	12.6
Standard Deviation	66,337.68		Standard Deviation	66,337.68	
Median	4,659.05		Median	4,659.05	
Observations	2,465		Observations	2,465	

Note: MV_t is defined as market value of equity in time t. MVWA (Apr) is measured by market value weighted average return of hedge portfolio for the year beginning April in year $t+1$.

top quintile of hedge portfolios over the other stocks hedge portfolios is 42.94% [= 42.25%−(−0.69%)] annually over the past 10 years. The empirical result shows the annual average return of the hedge portfolio is over 40% by standing on the regression analysis residuals one year ahead, which is much higher than the average return (12.6%) of all samples. Meanwhile, the average return of hedge portfolio formed with the ranked error terms is negatively related to error terms. This means the portfolios with smaller error terms (e.g., Quintile 1) gain higher average return, which is consistent with our expectation. We also suggest these seven financial performance measures can strongly explain not only the stock prices of those high-tech companies, but also help to predict future returns in Taiwan.

Further analyzing by year, we also find the return of hedge portfolios during the forecast period shows seven of ten years are significantly positive, and only one year is significantly negative (result from sensitivity analysis is not listed here).[7] However, as the influence of foreign capital increases and stock market matures, the returns of hedge portfolios become relatively smaller than ever.

Panel B of Table 4 shows the results of the tercile hedge portfolios' return-predictability performance and also find Tercile 1, with lower ranked error terms, still has the highest market value weighted average return; in contrast, Tercile 3, with higher ranked error terms, has the lowest market value weighted average return.

In sum, the results support our hypothesis and imply firms' current year estimation errors are negatively associated with subsequent stock returns.

4.3. Sensitivity Analysis

To check the robustness of our results, we use both stock prices at the end of June and end of December, and further recompute portfolio annual return based on the weighted average of market value (market value weighted average) and weighted average sum of firms (simple average) respectively in sensitivity analysis.

Panels A and B of Table 5 show the quintile and tercile hedge portfolios' return-predictability performance and reports the estimation results of the portfolios formed based on simple average and market value weighted average, respectively. From the results in Table 5, we have consistent conclusions as earlier stated that support our hypothesis.

Table 5. Estimation Results of Hedge Portfolios with Respect to Future Annual Stock Returns Sample Consists of 2,465 Firm-Years between 1995 and 2004.

Portfolios	MV_t	Simple Average			Market Value Weighted Average	
		R_{t+1} (Apr) (%)	R_{t+1} (Jun) (%)	R_{t+1} (Dec) (%)	MVWA (Jun) (%)	MVWA (Dec) (%)
Panel A: 5 hedge portfolios						
Quintile 1 (lowest error terms)						
Mean	15,639.72	27.12	22.80	27.89	52.95	38.34
Standard Deviation	38,126.91	58.43	48.49	60.14		
Median	5,851.20	10.76	9.92	10.62		
Observations	493					
Quintile 2						
Mean	8,523.50	41.69	32.76	27.93	24.38	23.20
Standard Deviation	16,192.59	87.93	73.58	68.76		
Median	3,830.10	16.58	18.64	9.28		
Observations	493					
Quintile 3						
Mean	12,967.22	22.35	19.78	19.60	16.49	18.57
Standard Deviation	25,321.97	52.87	54.99	52.70		
Median	5,138.70	9.26	6.65	5.13		
Observations	493					
Quintile 4						
Mean	19,491.02	31.70	29.68	19.95	17.92	13.65
Standard Deviation	50,905.82	70.95	63.38	56.10		
Median	5,208.90	11.52	16.99	5.93		
Observations	493					
Quintile 5 (highest error terms)						
Mean	44,572.83	13.49	4.90	12.65	4.07	2.66
Standard Deviation	113,556.98	47.44	44.63	54.33		
Median	7,586.30	5.16	−5.32	2.19		
Observations	493					
Full sample						
Mean	20,267.32	27.25	21.93	21.61	17.05	13.99
Standard Deviation	66,337.68	71.97	64.17	60.77		
Median	4,659.05	9.40	10.43	6.61		
Observations	2,465					
Panel B: 5 hedge portfolios						
Tercile 1 (lowest error terms)						
Mean	13,216.84	34.33	28.38	28.24	43.84	32.92
Standard Deviation	34,925.24	77.84	63.20	66.69		
Median	4,816.55	14.83	12.82	10.24		
Observations	822					

Table 5. (*Continued*)

Portfolios	MV_t	Simple Average			Market Value Weighted Average	
		R_{t+1} (Apr) (%)	R_{t+1} (Jun) (%)	R_{t+1} (Dec) (%)	MVWA (Jun) (%)	MVWA (Dec) (%)
Tercile 2						
Mean	11,894.26	29.67	24.94	22.26	13.09	17.71
Standard Deviation	28,110.95	68.73	61.12	58.00		
Median	4,053.50	8.42	15.89	6.51		
Observations	822					
Tercile 3 (highest error terms)						
Mean	35,634.03	21.37	14.38	14.95	8.72	5.69
Standard Deviation	95,572.92	58.15	57.03	53.64		
Median	8,384.00	9.09	6.20	1.82		
Observations	821					
Full sample						
Mean	20,267.32	27.25	21.93	21.61	17.05	13.99
Standard Deviation	66,337.68	71.97	64.17	60.77		
Median	4,659.05	9.40	10.43	6.61		
Observations	2,465					

Notes: MV_t is the market value of equity in year t. R_{t+1} (Apr) is measured by average annual stock return of hedge portfolio for the year beginning April in year t+1. R_{t+1} (Jun) is measured by average stock return of hedge portfolio for the period from beginning June in year t. R_{t+1} (Dec) is measured by average stock return of hedge portfolio for the period from beginning December in year t. MVWA (Apr) is measured by market value weighted average return of hedge portfolio for the year beginning April in year t+1. MVWA (Jun) is measured by market value weighted average return of hedge portfolio for the year beginning June in year t+1. MVWA (Dec) is measured by market value weighted average return of hedge portfolio for the period from beginning December in year t.

5. CONCLUSION AND DISCUSSION

The objective of fundamental analysis studies is to use accounting numbers to evaluate the firm (Penman, 2001; Frankel & Lee, 1998). Equity market value reflects the present value of investors' expected future cash flows or earnings. In related studies, researchers search for variables that can explain current stock prices or help to predict future firm value. Furthermore, several empirical studies document intrinsic values estimated using the residual income model can predict future returns (Lee, 1999). However, Kothari (2001) indicates the residual income model provides little guidance

in terms of why we should expect to predict future returns using estimated intrinsic values.

This study extends and aims to apply fundamental analysis based on seven financial measures to predict high-tech stock returns, which can help investors make investment decisions. The financial measures included in our modified residual income model are the major components of a firm's financial statement. We use book value as a basic firm value measure and decompose cash flows into CFO, CFI, and CFF, and net income into R&D, OI, and ANOI as additional performance measures. Including these seven financial statement items in the valuation model can properly explain current firm value and help us predict future stock return through using the estimation errors to indicate the direction of one-year-ahead price changes.

For a sample of Taiwan information electronic firms, hedge portfolios involved in the trading strategy are operated on the basis of ranked estimation errors in long and short positions. The highest average one-year holding-period return of the hedge portfolios is about 43%, much higher than the risk free rate and average stock return (12.6%) in the same period.

Thus, we modify the residual income model and find the results of this simpler trading strategy approach can support the conclusion for the predictability of R&D, OI, ANOI, CFI, CFF, and CFO in future stock returns. The results also demonstrate the estimation error can be a good stock return predictor, but the return of hedge portfolios generally decreases as the market matures.

The results of this study deviate from the efficient market's view and show stock prices do not fully reflect all publicly available information. Our findings contribute to the finance and fundamental analysis literatures on the predictability of stock returns by documenting the return predictability of seven major financial performance measures, especially for the informational electronic industry, which is high-tech and highly competitive.

Moreover, by recognizing the economic effect of past transactions and events, past transactions have predictive ability for future events that the financial statements convey valuable information about firm's future value. The components of cash flows and various earnings-related items really provide significant explanatory power in relation to a firm's market value. Thus, just as the FASB advocates, present and potential investors can rely on the accounting information in valuing the firm to improve investment decisions.

We recognize the limitations of this study that (1) cross-sectional return predictability tests of market efficiency cannot invariably examine long-horizon returns; (2) the determinants of expected return are likely to be correlated with the portfolio formation procedure; (3) the survival bias and data problems may be serious; and (4) the risk of the hedge portfolio could not be reduced to zero because of some restrictions on short sales in the Taiwan stock market.

NOTES

1. Fama (1991) indicates returns can be predictable from past returns, dividend yields, and various term-structure variables. Beaver (2002) also asserts capital markets are inefficient with regard to at least three regions: postearnings announcement drift, market-to-book ratios and its refinements, and contextual accounting issues.
2. Sloan (1996) suggests an analysis of this type can be used to detect mispriced securities. Moreover, the related studies such as Frankel and Lee (1998), Dechow et al. (1999), and Lee et al. (1999), use the residual income model combined with analysts' forecasts to estimate fundamental values and shows abnormal returns can be earned. However, due to lack of reliable analysts' forecasts data in Taiwan, this study adopts the components of cash flow and earnings to predict future stock prices.
3. Major components of balance sheet and income statement are used instead of aggregating book equity and net income to avoid the severe inferential distortions that can arise when evaluating the value relevance of financial statements of fast growing, highly intangible-intensive companies (Zhang, 2001; Hand, 2004, 2005).
4. Because all listed companies in Taiwan announce their annual report and financial statements before the end of April; this study adopts the stock prices of end of April as the actual annual stock price.
5. The hedged portfolios in our trading strategy are formed annually by assigning firms into quintiles based on the magnitude of the absolute value of error terms in year t. Then equal-weighted stock returns are computed for each quintile portfolio over the subsequent year, beginning four months after the end of the fiscal year from which the historical forecast data are obtained.
6. In Taiwan stock market, the risk of this hedge portfolio could not be entirely eliminated because of some restrictions on short sale of securities. Therefore, the investors still bear some uncertain risk even if they operate the trading strategy through the short sale of Electronic Sector Index Futures.
7. The hedged portfolio return summarizes the predictive ability of our model with respect to future returns. Related statistical inference is conducted using the standard error of the annual mean hedged portfolio returns over the 10 years in the sample period.

REFERENCES

Abarbanell, J., & Bushee, B. (1997). Fundamental analysis, future earnings, and stock prices. *Journal of Accounting Research*, *35*, 1–24.

Abarbanell, J., & Bushee, B. (1998). Abnormal returns to a fundamental analysis strategy. *The Accounting Review*, *73*, 19–45.

Ball, R., & Brown, P. (1968). An empirical evaluation of accounting income numbers. *Journal of Accounting Research*, *6*, 159–177.

Beaver, W. H. (1968). The information content of annual earnings announcements. *Journal of Accounting Research*, *6*, 67–92.

Beaver, W. H. (2002). Perspectives on recent capital market research. *The Accounting Review*, *77*, 453–474.

Beneish, M. D., Lee, M. C., & Tarpley, R. L. (2001). Contextual fundamental analysis through the prediction of extreme returns. *Review of Accounting Studies*, *6*, 165–189.

Bernard, V., & Stober, T. (1989). The nature and amount of information in cash flows and accruals. *The Accounting Review*, *64*, 624–652.

Bernard, V., & Thomas, J. (1989). Post-earnings announcement drift: Delayed price response or risk premium? *Journal of Accounting Research*, *27*, 1–36.

Bernard, V., & Thomas, J. (1990). Evidence that stock prices do not fully reflect the implications of current earnings for future earnings. *Journal of Accounting and Economics*, *13*, 305–340.

Black, E. L. (1998). Life-cycle impacts on the incremental value-relevance of earnings and cash flow measures. *Journal of Financial Statement Analysis*, *4*, 40–56.

Bowen, R., Burgstahler, D., & Daley, L. (1986). Evidence on the relationships between earnings and various measures of cash flow. *The Accounting Review*, *61*, 713–725.

Bowen, R., Burgstahler, D., & Daley, L. (1987). The incremental information content of accrual versus cash flows. *The Accounting Review*, *62*, 723–747.

Chan, K. C., Lakonishok, J., & Sougiannis, T. (1999). *The stock market valuation of research and development expenditures*. Working Paper. University of Illinois.

Chen, P., & Zhang, G. (2007). How do accounting variables explain stock price movements? Theory and evidence. *Journal of Accounting and Economics*, *43*, 219–244.

Dechow, P. (1994). Accounting earnings and cash flows as measures of firm performance: The role of accounting accruals. *Journal of Accounting and Economics*, *18*, 3–42.

Dechow, P., Hutton, A., & Sloan, R. (1999). An empirical assessment of the residual income valuation model. *Journal of Accounting and Economics*, *26*, 1–34.

Fairfield, P. M., Sweeney, R. J., & Yohn, T. L. (1996). Accounting classification and the predictive content of earnings. *The Accounting Review*, *71*, 337–355.

Fama, E. F. (1991). Efficient capital market: II. *Journal of Finance*, *46*, 1575–1617.

Feltham, G., & Ohlson, J. (1995). Valuation and clean surplus accounting for operating and financial activities. *Contemporary Accounting Research*, *11*, 689–731.

Frankel, R., & Lee, C. (1998). Accounting valuation, market expectation, and cross-sectional stock returns. *Journal of Accounting and Economics*, *25*, 283–319.

Fu, C. J., & Chang, Y. L. (2006). *The impacts of life cycle on the value relevance of financial performance measures: Evidence from Taiwan*. Working Paper. 2006 Annual Meeting of the American Accounting Association.

Grossman, S. J., & Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *American Economic Review*, *70*, 393–408.

Hand, J. R. M. (1990). A test of the extended functional fixation hypothesis. *The Accounting Review*, *65*, 740–763.

Hand, J. R. M. (2003). *The relevance of financial statements within and across private and public equity markets*. Working Paper. UNC Chapel Hill: Kenan-Flagler Business School.

Hand, J. R. M. (2004). The market valuation of biotechnology firms and biotechnology R&D. In: J. McCahery & L. Renneboog (Eds), *Venture capital contraction and the valuation of high-technology firms* (pp. 251–280). UK: Oxford University Press.

Hand, J. R. M. (2005). The value relevance of financial statements in the venture capital market. *The Accounting Review*, *80*, 613–648.

Holthausen, R., & Larcker, D. (1992). The prediction of stock returns using financial statement information. *Journal of Accounting and Economics*, *15*, 373–411.

Ittner, C. D., Larcker, D. F., & Randall, T. (2003). Performance implications of strategic performance measurement in financial service firms. *Accounting, Organizations and Society*, *28*, 715–741.

Jorion, P., & Talmor, E. (2001). *Value relevance of financial and nonfinancial information in emerging industries: The changing role of web traffic data*. Working Paper. University of California at Irvine.

Klein, A., & Marquardt, C. (2006). Fundamentals of accounting losses. *The Accounting Review*, *81*, 179–206.

Kothari, S. P. (2001). Capital markets research in accounting. *Journal of Accounting and Economics*, *31*, 105–231.

Lee, C. (1999). Accounting-based valuation: Impact on business practices and research. *Accounting Horizons*, *13*, 413–425.

Lee, C., Myers, J., & Swaminathan, B. (1999). What is the intrinsic value of the Dow? *Journal of Finance*, *54*, 1693–1741.

Lev, B., & Thiagarajan, R. (1993). Fundamental information analysis. *Journal of Accounting Research*, *31*, 190–215.

Maines, L. A., & Hand, J. R. (1996). Individuals' perceptions and misperceptions of time series properties of quarterly earnings. *The Accounting Review*, *71*, 317–336.

Ohlson, J. (1995). Earnings, book values and dividends in security valuation. *Contemporary Accounting Research*, *11*, 661–687.

Ou, J., & Penman, S. (1989a). Financial statement analysis and the prediction of stock returns. *Journal of Accounting and Economics*, *11*, 295–329.

Ou, J., & Penman, S. (1989b). Accounting measurement, price–earnings ratios, and the information content of security prices. *Journal of Accounting Research*, *27*, 111–152.

Penman, S. H. (2001). *Financial statement analysis and security valuation* (pp. 2–20). Irwin, NY: McGraw-Hill.

Penman, S. H., & Sougiannis, T. (1998). A comparison of dividend, cash flow, and earnings approaches to equity valuation. *Contemporary Accounting Research*, *15*, 343–383.

Rayburn, J. (1986). The association of operating cash flow and accruals with security returns. *Journal of Accounting Research*, *24*, 112–133.

Robinson, K. C. (1998). An examination of the influence of industry structure on the eight alternative measures of new venture performance for high potential independent new ventures. *Journal of Business Venturing*, *14*, 165–187.

Sloan, R. (1996). Do stock prices fully reflect information in accruals and cash flows about future earnings. *The Accounting Review*, *71*, 289–315.

Zhang, X.-J. (2001). *Accounting conservatism and the analysis of line items in earnings forecasting and equity valuation*. Working Paper. University of California, Berkeley.

This page intentionally left blank

FORECASTING INFORMED TRADING AT MERGER ANNOUNCEMENTS: THE USE OF LIQUIDITY TRADING

Rebecca Abraham and Charles Harrington

ABSTRACT

We propose a novel method of forecasting the level of informed trading at merger announcements. Informed traders typically take advantage of their knowledge of the forthcoming merger by trading heavily at announcement. They trade on positive volume or informed buys for cash mergers and negative volume or informed sells for stock mergers. In response, market makers set wider spreads and raise prices for informed buys and lower prices for informed sells. As liquidity traders trade on these prices, our vector autoregressive framework establishes the link between informed trading and liquidity trading through price changes. As long as the link holds, informed trading may be detected by measuring levels of liquidity trading. We observe the link during the −1 to +1 period for cash mergers and −1 to +5 period for stock mergers.

Advances in Business and Management Forecasting, Volume 6, 37–51

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006003

INTRODUCTION

The financial markets typically contain traders with different motivations and trading strategies. Informed traders act on the basis of privileged information. At earnings announcements, certain firms report earnings surprises, announcing earnings that are either higher or lower than analysts' forecasts. Dividend announcements involve the announcement that the firm is about to pay dividends or increase its dividend payout ratio. At merger announcements, target firms are perceived as valuable in that they are desired by acquirers. Firms facing bankruptcies send negative signals into the markets. Informed traders capitalize on their knowledge of the positive signals generated by positive earnings surprises, dividend announcements, and rises in target firms' prices by buying stock or call options, and selling stock or buying put options on negative signals such as negative earnings surprises or forthcoming bankruptcies. In each case, the informed trader is aware of the firms' performance and is capable of profiting from that information. By definition, uninformed trading is termed noise trading. Black (1986) offers the following definition:

> Noise trading is trading on noise as if it were information. People who trade on noise are willing to trade even though from an objective point of view they would be better off not trading. (Black, 1986, p. 529)

Uninformed traders, on the other hand, trade at random. They merely buy and sell for their own inventory. Uninformed traders, also termed liquidity traders, are driven by the desire to maintain liquidity. Therefore, at mergers, earnings announcements, or dividend announcements, liquidity traders trade to maintain an inventory of stock, or in response to pressure from their clients (liquidity traders are typically market makers and institutional traders). They are not motivated by the profit-making desires of informed traders. If they take large positions, it is usually after information about the event (merger or dividend announcement) has become public with a view to satisfying their clients who wish to hold target stock or stock on which dividends have just been announced.

The literature is replete with reports of trading by informed traders (Admati & Pfleiderer, 1988; Glosten & Milgrom, 1985; Kyle, 1985; Lee & Yi, 2001) on different occasions. These papers also develop theoretical constructs by which the impact of informed traders on the markets may be derived. Market makers in the stock market buy and sell stock to traders at bid and ask prices, respectively. If a trader wishes to purchase, the market maker offers a variety of bid prices; likewise, if another trader wishes to sell the

market maker offers a variety of ask prices. The difference between the bid and ask prices is termed the quoted spread. Setting of the spread is the mechanism by which market makers influence prices. For informed traders, the market maker, knowing the trader will profit at his (the market maker's) expense, will set the spreads to be wide to cover losses. In contrast, for liquidity traders, the market maker will set spreads to be narrow as losses are not expected to occur. It is the position of this chapter that the influence of spreads (based on informed trading volumes) on prices will determine the volume of liquidity trading. Accordingly, we focus on establishing a link between volumes of informed trading, price changes resulting from spread widths, and liquidity trading as a means of forecasting the level of informed trading.

Why is it important to forecast levels of informed trading volume? Regulators in recent years have embarked on a campaign to reduce spreads. High spreads contribute to higher transactions prices for traders. Higher transactions prices suggest less market efficiency, higher costs to brokerage houses and their clients for purchasing stock or options, and monopoly profits for market makers. Consequently, regulators have ordered the listing of options on multiple exchanges so that competition between market makers will drive down spreads and reduce transaction fees for options traders. At present, almost 80% of all traded options are listed on all six options exchanges (De Fontnouvelle, Fishe, & Harris, 2003). This outcome has been achieved after passage of Rule 18-c by the Securities and Exchange Commission in 1981, banning uncompetitive practices, a lawsuit and settlement by the Department of Justice against the American Stock Exchange in 1989, and an options campaign on August–September 1989, designed to increase multiple listing. As informed traders contribute to higher spreads, the ability to forecast levels of informed trading is of practical interest to regulators tracking the incidence of rising spreads. From a theoretical standpoint, most models of informed trading (Admati & Pfleiderer, 1988l; Bamber, Barron, & Stober, 1999) have viewed informed and liquidity trading as tangentially linked. It would be useful to establish a direct link between these two trading activities. It is to that purpose that this study is directed.

REVIEW OF LITERATURE

This chapter casts informed and liquidity trading in the context of merger announcements. Merger announcements have been studied extensively as a venue of information-based trading (for a review, see Cao, Chen, & Griffin, 2005). However, the focus of merger research has been on the target firm.

Target firms rise in price with the announcement of a merger as the signal is sent that they are in demand by acquirer firms. We wish to shift the focus to acquirer firms. The cornerstone of our work is the response by informed traders to signals. Therefore, the signal should be as pure and unidirectional as possible. If all mergers are considered as similar, there may be mixed signals whose effects obliterate each other. One method of obtaining pure signals is by classifying mergers in terms of method of payment. Mitchell, Pulvino, and Stafford (2004) conceived of mergers as cash mergers if the acquirer paid cash for the acquisition or a stock (fixed-exchange ratio) merger if stock was exchanged in a fixed ratio at the time of closing. Cash mergers are viewed as positive in that the acquiring firm must have excess cash to spend on a purchase. Conversely, acquirers that engage in a stock exchange are viewed unfavorably as being devoid of financial resources. At merger announcement, Mitchell et al. (2004) observed significant positive cumulative abnormal returns (CAARs) on cash mergers and significant negative CAARs on stock mergers. Intuitively, informed traders would use their privileged information of a positive signal on forthcoming cash mergers to purchase acquirer stock, or we would expect informed buying volume to rise at the time of announcement of a cash merger. In contrast, informed traders would use their privileged information of the negative signal on stock mergers to sell acquirer stock at merger announcements. Therefore, at the time of merger announcement, typically, days -1 to $+1$ of the merger, the volume of trading on both cash and stock acquirers may be expected to rise significantly above normal trading volume.

H1a. CAARs due to elevated total volume of trading on cash acquirers will be significantly positive during days -1 to $+1$ of a merger announcement than during the benchmark period.

H1b. CAARs due to elevated total volume of trading on stock acquirers will be significantly negative during days -1 to $+1$ of a merger announcement than during the benchmark period.

Theoretical Framework

Easley and O'Hara (1992) propose a model of informed trading and liquidity trading that may be used as the basis for this study. They view informed traders as arriving in a continuous auction. Informed traders trade on the same side of the market. For cash mergers, informed traders will buy

only while for stock mergers they will only sell. Informed traders trade in large sizes (Easley & O'Hara, 1992) as they wish to make as large a profit as possible with a single trade. In fact, the large size characteristic of informed trades led Heflin and Shaw (2005) to develop a method of detecting informed trades by transaction. Informed traders also trade regularly until prices have reached a market equilibrium. Market makers, on the other hand, are in an adversarial position vis a vis informed traders. The increased trading of informed traders suggests larger losses to them as does large size trades. As far as the market maker is concerned, informed traders trading in block trades reduce their profits substantially. In response, market makers will increase quoted spreads (ask price − bid price), or charge higher prices to informed traders who wish to buy (as in cash mergers) and pay lower prices to informed traders who wish to sell (as in stock mergers). This results in higher effective spreads (trade price − [bid price + ask price]/2), which are the more relevant measures as they include the actual price at which trading took place. Therefore, trade prices will rise on buy orders for cash mergers and fall for sell orders with stock mergers. Information is thus impounded into prices.

However, informed traders are not the only traders in the market. The market maker is aware of the existence of uninformed or liquidity traders. As informed traders continue to trade, market makers will, over time, try to infer the percentage of trading volume that is informed versus uninformed. The composition of trades may provide some information to the market maker. As the informed all trade on the same side of the market, the percentage of such trades will indicate the proportion of informed trading volume. For cash mergers, all informed traders may be expected to buy. Uninformed traders will both buy and sell as they are purchasing and selling at random. The buy trades may be attributed to informed and uninformed traders. As informed traders are assumed to trade in large size only, whereas some of the uninformed may trade in large size, all small trades are eliminated. The only doubt remains on large trades, which may belong to both informed or uninformed traders. Market makers can infer the existence of informed traders by identifying large buy trades for cash mergers and large sell trades for stock mergers. Over time, with repeated purchases by informed traders, market makers are able to identify the trades that are informed. Accordingly, they will adjust spreads downward for trades that are clearly uninformed trades, while maintaining high spreads for informed trades. Soon after this occurs, information will no longer enter into prices and prices will clear at an equilibrium level. This suggests that a multiperiod model is more appropriate, in that spread and price adjustments cannot take place instantaneously. It is possible that adjustments may take place on an

intraday basis or over the span of several days. In either case, empirical testing must use lagged structures to capture intraday price changes and spread analysis over several days to explore interday effects.

What is the reaction of liquidity traders? Liquidity traders are not novices. In fact, most of them are institutional traders. They just have different motivations for trading from the informed traders. They choose to purchase and sell for reasons of ownership rather than acting to profit on the basis of specialized information. They are aware of spreads and transactions costs. In fact, the Easley and O'Hara (1992) model makes the assumption that there is a constant (nonincreasing) transaction cost that liquidity traders assume as the model could not exist without liquidity traders. Liquidity traders accept and pay higher prices for cash buys and earn lower amounts on sales in stock mergers. The question may be raised as to why they accept these conditions? Why do liquidity traders accept the higher spreads and unfavorable prices which are the market makers' reaction to informed trading? Why do not they leave the market (Black, 1986). Trueman (1988) advances the explanation that investors view fund managers as being more productive, and in turn invest in a fund if the total volume of trading in the fund is large. To the investor, a large volume of trading suggests more information-based trades and thereby higher fund profits. However, the investor does not know the proportion of trades that are information-based; he or she simply views the total amount of trading and assumes that a larger trading volume indicates more information-based trades. This provides an incentive for fund managers to engage in liquidity trading to send the signal to the investor that he or she is actually making information-based trades, and thereby attract investor funds. Empirically, Kanodia, Bushman, and Dickhaut (1986) observed that managers do not forego unprofitable investment projects as withdrawal would send the message that the manager's information about the outcome of the project was inaccurate. Trueman (1988) showed that security analysts may be unwilling to revise original erroneous estimates on receiving new information due to the negative signal such action would send about the wisdom of the original forecast.

For our purposes, it is apparent that liquidity traders will continue to trade even at the unfavorable prices set by market makers in response to informed trading. The volume of informed trading influences stock price changes to be either positive (for cash mergers) or negative (for stock mergers). Those price changes will influence the level of liquidity trading. Assuming that liquidity traders will continue to trade, i.e., that they continue to trade to attract future investment, they will pay higher prices for cash mergers and lower prices for stock mergers. The trading volume of liquidity traders will depend

on the level of trading of informed traders. There is a three-step link with higher informed buying (selling) resulting in higher (lower) stock price changes and in turn, higher liquidity trading as liquidity traders accept the adverse prices set by market makers responding to informed trading. It is this link that this study tests, and which acts as a forecast of informed trading using liquidity trading volumes. This link maybe expected to weaken over time, as market makers learn the pattern of informed trading and set differential spreads for informed and liquidity traders. Liquidity trading at that point will no longer depend on informed trading and it will be no longer possible to forecast the existence of informed trading using liquidity trades.

H2a. Liquidity buying volume for acquirer stock in cash mergers will be directly related to informed buying volume and stock price changes. Specifically, informed buying volume will increase future stock price changes, which will increase liquidity buying volume on days -1 to $+1$ of the merger announcement.

H2b. Liquidity selling volume for acquirer stock in stock mergers will be directly related to informed selling volume and stock price changes. Specifically, informed selling volume will decrease future stock price changes, which will reduce liquidity buying volume on days -1 to $+1$ of the merger announcement.

DATA AND SAMPLE CHARACTERISTICS

All cash and stock mergers for 2005 were obtained from Factset Mergerstat's comprehensive database of worldwide merger transactions. Mergers were filtered by completion, as only acquisitions of complete firms were included. Partial transfers including sales of divisions and assets were not included. All foreign firms (either acquirer or target) were excluded. Although data is available for the past 25 years, only 1 year of data was included as the vector autoregressions used to test Hypothesis 2 use intraday data. Intraday data is extremely voluminous, with >10 million data points for the sample period. All mergers were examined closely for method of payment, with only full cash purchases and fixed ratios of exchanges of stock between target and acquirer being included. The final sample consisted of 66 cash mergers and 38 stock mergers. Intraday stock trade and quote prices were obtained from the New York Stock Exchange's TAQ (Trade and Quotes) database. TAQ provided intraday bid and ask quotes for each trade

occurring on each stock in the database from 9:30 a.m. to 4:00 p.m. Trade prices and offer quantities (offer size) were provided as well.

METHODOLOGY

The complete sample of cash and stock mergers were subjected to a volume event study. Volume event studies are based on the market or capital asset pricing model (CAPM). CAPM is represented by the following equation:

$$E(r_j) = R_f + \beta(R_m - R_f) \tag{1}$$

The expected return on a stock is a function of the stock's excess return above the risk-free rate weighted by the stock's correlation with the market or the β coefficient. Volume event studies find the cumulative excess return over a period due to abnormal volumes over a benchmark or "normal" period usually of 255-day duration prior to the model period. Volume event studies are particularly useful for our purposes as they can measure the extent of excessive total trading volume during the merger period. Separate volume event studies for cash and stock mergers were used to determine the period to be considered in future analysis; even though the merger announcement period is generally taken to be days -1 to $+1$ (Mitchell et al., 2004), it is possible for the year in our sample to have a slightly different (wider or narrower) announcement period. As the volume of trading was measured over several days, the two volume event studies tested Hypothesis 1. The Center for Research in Security Prices (CRSP) provided the risk-free rate, market return, and β coefficient, and daily stock closing stock volume inputs into the event studies. The Eventus program used the CRSP inputs to compute CAARs reported in Table 1.

Hypothesis 2 was tested by a vector autoregressive model. Vector autoregressions are employed when it is necessary to link multiple variables as in this study which seeks to establish a path from informed trading volume to stock price changes to liquidity trading volume for two different types of mergers over several days. The other condition for usage of vector autoregressions as opposed to two-stage least squares is the unknown lag structure of the informed buy and liquidity buy volumes.

The following relationships were tested:

$$R_t = a_1 r_{t-1} + a_2 r_{t-2} + \cdots + a_n \mathrm{pr}_{t-n} + b_0 s_t + b_1 s_{t-1} + \cdots + b_n s_{t-n} \tag{2}$$

$$U_t = a_1 r_{t-1} + a_2 r_{t-2} + \cdots + a_n \mathrm{pr}_{t-n} + b_0 u_t + b_1 u_{t-1} + \cdots + b_n u_{t-n} \tag{3}$$

Table 1. Results of Volume Event Studies of Cash and Stock Mergers.

Cash Mergers			Stock Mergers		
Event window	CAAR	Patell Z	Event window	CAAR	Patell Z
−1, 0	42.02%	1.927*	−1, 0	247.95%	25.953***
Day	CAAR	Patell Z	Day	CAAR	Patell Z
−1	−20.45%	0.043	−1	−27.75%	0.455
0	62.46	2.682***	0	275.70	36.249*
1	8.42	5.936***	1	86.97	17.873***
			2	14.17	6.211***
			3	−16.32	4.77***
			4	−11.79	3.772***
			5	−0.64	3.776***

*$p < .01$; ***$p < .001$.

where R_t is the contemporaneous and lagged stock price changes; s_t the informed buy volume for cash mergers and informed sell volume for stock mergers at time t; and U_t the liquidity buy volume for cash mergers and liquidity sell volume for stock mergers at time t.

Input variables into the vector autoregression included an intraday stock price change series, informed trading volume, and liquidity trading volume. The stock price change series was constructed by finding the mean stock prices for each 5-min interval from 9:30 a.m. to 4:00 p.m. for each day in the merger announcement period. The difference between consecutive mean prices was obtained to get a change series. Z scores were computed by finding the difference between each stock price change and mean daily changes and dividing by the standard deviation of daily stock price changes. This procedure was considered necessary to smooth out the autocorrelation of daily stock price change data. Heflin and Shaw's (2005) methodology was used to differentiate between informed and liquidity trades. They argued that a trade size/offer quantity ratio <1 indicated an informed trade, whereas that above 1 indicated a liquidity trade. Their rationale was that informed traders prefer to trade in large size, but market makers will seek to limit the size of informed trade so that the market makers will only permit trades with sizes below the offer quantity to be completed. For liquidity trades, market makers will permit large trades, which are above the offer quantity. Accordingly, we computed trade size/offer quantity ratios for all trades for each day in the announcement period, differentiating them as

informed or liquidity. The Lee and Ready algorithm was employed to determine whether trading volumes were buy or sell volumes. For each informed or liquidity trade, if the trade size > bid ask midpoint, the trade was designated an informed buy trade or liquidity buy trade. Likewise, if the trade size < bid ask midpoint, the trade was designated an informed or liquidity sell trade. The volumes for each trade were consolidated into 5-min intervals to match the stock price change series. To smooth out fluctuations in raw data, trade volumes were converted into Z scores by subtracting mean daily trade volumes and dividing by the daily standard deviation of trade volumes.

RESULTS

The entire cash merger sample of 66 events was subjected to a volume event study. Twelve events were dropped by Eventus, leaving a final sample of 54 events. Table 1 reports the results of this event study. The day −1 to day +1 merger announcement period window showed a highly significant CAAR of 50.01% (day −1 to 0, 42.2%, Patell $Z = 1.927$, $p < .05$). Excess volume was confined to the announcement period and day +2, with significant daily CAARs for day 0 of 62.46% (Patell $Z = 2.682$, $p < .01$), day 1 of 8.42% (Patell $Z = 5.936$, $p < .001$), and day 2 of −14.22% (Patell $Z = 3.104$, $p < .001$). The shift in volumes from positive to negative on day +2 indicates the beginning of price reversal, or that the entire trading on the positive signal due to the cash merger is absorbed into prices at the end of day +1. Hypothesis 1a was supported for cash mergers; a significant positive CAAR was found for each day in the announcement period. For the stock mergers, two events were dropped by Eventus, leaving a final sample of 36 events. Table 1 reports the results of this event study. Hypothesis 1b was partially supported. Significant negative CAARs were obtained for days 3 and 4, (−16.32% and −11.79% with Patell Z values of −4.77, $p < .001$, and 3.772, $p < .001$, respectively). Days 0, 1, and +2 showed highly significant positive CAARs of 275.70% on day 0 (Patell $Z = 36.249$, $p < .001$), 86.97% on day 1 (Patell $Z = 17.873$, $p < .001$), and 14.17% (Patell $Z = 6.211$, $p < .001$). The change in sign may be due to both informed buyers and sellers being active in the market.

Table 2 reports descriptive statistics for liquidity buyers and sellers throughout the announcement period for both samples. The number of trades, mean, standard deviation, and maximum and minimum trade sizes are reported.

Table 2. Descriptive Statistics for Liquidity Trading Volumes, Cash and Stock Mergers on Acquirer Stock (2005).

Day	Range of trades	Mean trade size	SD trade size	Maximum size	Minimum size
Cash Mergers		Informed Buy Volume			
−1	124–2,429	2,273.766	3,836.35	500,000	101
0	216–3,379	2,312.493	2,909.312	500,000	101
1	19–3,172	2,507.624	2,786.744	2,000,000	102
Stock Mergers		Informed Sell Volume			
−1	45–1,680	1,255.278	646.996	143,100	101
0	1,320–3,615	2,519.937	2,569.231	2,013,100	101
1	45–1,406	1,820.686	1,489.293	248,600	101
2	9–883	1,539.899	1,416.966	250,000	101
3	35–1,457	1,615.354	1,142.617	338,800	101
4	41–779	1,629.203	1,685.792	212,000	101
5	45–652	1,923.973	2,268.493	900,000	101

Table 3 shows the results of the vector autoregressive model for cash mergers relating informed buy volume to stock price changes to liquidity buy volume. On day −1, informed buy volume significantly increased stock price changes in the third lag. Buying by informed traders led to higher stock price changes 15 min later (as each lag represents a 5-min interval), which in turn led to buying by liquidity traders in lags 1–6 or 5–30 min later. On day 0, adjustments of stock prices to information proceeds much faster, presumably because it is the day of announcement. Informed buying volume significantly influences stock prices 5 min later, which in turn raise liquidity buying volume in the first lag or 5 min later. Therefore, Hypothesis 2a is supported for days −1 and 0. It is not supported for day 1. On day 1, the stock price changes significantly increase liquidity buy volume in the first lag, or 5 min later. However, informed buying volume has not have any effect on stock price changes, possibly because market makers have learned to identify informed trades, so that an informed trading strategy no longer guarantees excess profits. For stock mergers, Hypothesis 2b was supported for day −1 as informed selling significantly influenced stock price changes in the third and eighth lags, whereas stock price changes significantly predicted liquidity sales volume in the first lag. Selling by informed traders lowered stock prices about 15 min later, which in turn induced liquidity

Table 3. Results of Vector Autoregressions of Liquidity Volume on Stock Price Changes and Informed Volume.

Day −1	Coefficients	Day 0	Coefficients	Day 1	Coefficients
Stock price changes	Buy volume	Stock price delta	Buy volume	Stock price delta	Buy volume
Cash mergers					
0.0857***	0.0345	0.0941*	0.1086*	0.0571***	−0.040
−0.0015	−0.0153	0.0127	0.0366	−0.035	−0.028
−0.0119	0.0622**	−0.002	0.0103	0.0163	−0.025
0.0156	−0.0333	−0.0016	0.0134	−0.0028	−0.030
0.0402	−0.0134	0.0251	0.0069	0.0195	−0.036
−0.0407	0.0220	−0.0204	−0.0126	−0.0253	−0.037
0.0253	0.0067	−0.0066	0.0233	−0.0242	−0.021
0.0304	−0.0079	−0.0237	−0.0252	0.0067	−0.027
−0.0491	−0.0185	0.0199	0.0050	−0.0163	−0.039
R^2 0.058		0.049		0.163	
N 2,097		2,696		2,721	
Stock price changes	Sell volume	Stock price delta	Sell volume	Stock price delta	Sell volume
Stock mergers					
−0.0918*	0.0844	−0.2978**	0.1744	−0.1999***	0.0275
0.084	0.0115	−0.1195	−0.0284	−0.0001	0.0569
−0.0423	−0.0523*	0.0875	−0.2734*	−0.0037	−0.0001
−0.0219	−0.0056	0.2195	0.0589	0.0104	0.0628
−0.0052	−0.0145	−0.1216	−0.2183	−0.0617	−0.0427
−0.0003	0.0193	−0.0905	0.0432	−0.0196	0.0279
−0.0338	0.0106	0.0879	−0.1363	0.0167	−0.0429
−0.0146	−0.0493*	0.0972	−0.0996	−0.0108	−0.0923*
0.0001	−0.0148	0.0849	−0.0493	0.0493	0.010
R^2 0.125		0.292		0.1286	
N 1,083		1,083		1,083	
Day −1	Coefficients	Day 0	Coefficients	Day 1	Coefficients
Stock mergers					
−0.0744*	−0.0867**	−0.1005***	0.1552***	−0.0085	0.071
0.0079	0.0350	−0.0185	0.0683	−0.0135	0.073
−0.0029	0.0593	−0.0756*	0.0001	−0.0308	0.010
0.0597	0.0495	0.0185	0.0501	0.0077	−0.001
0.0515	−0.0080	0.0038	−0.0355	−0.0013	−0.001
0.0081	0.0636	−0.0245	−0.0198	0.0075	−0.011
0.0008	0.0415	0.0684	0.0251	0.0461	−0.005
−0.0355	−0.0132	−0.0050	−0.0061	−0.0313	0.141
0.0178	−0.0111	−0.0040	0.0300	−0.0554	−0.036
R^2 0.032		0.073		0.047	
N 1,083		1,083		1,083	

Table 3. (*Continued*)

Day 5	Coefficients
Stock price changes	Sell volume
Stock mergers	
−0.1284	0.1700
0.0616	−0.0612
−0.0956	−0.1271
−0.2558*	0.1102
0.3274	0.0546
−0.0414	0.2540
−0.0460	−0.2075
0.1742	0.0151
−0.0217	−0.1638
$R^2 = 0.311$	
$N = 1{,}083$	

*$p < .05$; **$p < .01$; ***$p < .001$.

selling 5 min later. On day 0, lag patterns were identical to day −1. Hypothesis 2b was supported with informed selling significantly influencing stock price changes in the third lag and stock price changes predicting liquidity selling in the first lag. On day 1, Hypothesis 2b was supported although stock price changes responded more slowly to informed selling volume. Informed selling volume significantly decreased stock prices in the eighth lag only – a full 40 min after the initial signal. Stock price changes influenced liquidity selling rapidly in the first lag or 5 min. Market makers may have been deliberating about spreads and the information received over several days of informed trading. Spreads may have been lowered so that trading by informed traders may have started to have a more tenuous effect on security prices. Hypothesis 2b continued to be supported on days 2 and 3. On days 2 and 3, informed selling affected stock price changes within 5 min and in turn liquidity selling 10 min later. Price adjustments were rapid due to market makers finally being able to identify informed trades and therefore, adjusting spreads accordingly. It is possible that full knowledge of informed trading was not gleaned, as some informed trading did affect stock prices and liquidity trading. The link ended on days 4 and 5, with informed trading having no significant influence on stock price changes and liquidity selling, presumably as market makers were able to correctly distinguish between informed and liquidity trades and set spreads accurately. Therefore, Hypothesis 2b was not supported for days 4 and 5.

CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

This study has provided empirical proof of the Easley and O'Hara (1992) model. Informed buying and selling volumes act to increase or decrease stock price changes which affect liquidity trading. By examining the pattern of relationships defined by our vector autoregressions, it is possible to forecast the actions of informed traders over the merger announcement period. As long as the link between informed trading, stock price changes, and liquidity trading holds, market makers in the stock market at merger announcements are setting high spreads, raising bid and ask quotes, and impounding information into prices. In other words, market makers are responding to the actions of informed traders, or informed trading is driving prices and uninformed trades. This occurred throughout the -1 to $+1$ merger announcement period for cash mergers. For stock mergers, the situation was found to be inherently more complex. For days -1 to $+1$, informed traders influenced stock price changes and liquidity trading. On days 2 and 3, there was rapid inclusion of informed trading in stock prices as stock prices achieved equilibrium rapidly. The existence of some market imperfection was apparent in that some trading on privileged information did have an impact on security prices. However, it is evident that market makers were beginning to infer patterns of informed trading and distinguish them from liquidity trading. Spreads began to be adjusted downwards, and excessive profits were not as easily available to informed traders as on the previous days. On days 4 and 5, all excess profits disappeared as market makers finally adjusted spreads to equilibrium prices and liquidity traders were subject to transaction costs that reflected their true values. This study has provided the first evidence of trading patterns for informed and liquidity traders during merger announcements, where the type of merger or method of payment forms the basis of demarcation. We have found that spreads and prices do not adjust to equilibrium levels for cash mergers during the announcement period. For stock mergers, they adjust over a longer period, i.e., during the fourth and fifth days. Hypothesis 1b was partly supported for stock mergers with shifts from significant positive to negative CAARs during the day 0 to $+5$ period. It is possible that there are multiple groups of informed or liquidity traders. Future autoregressive models should include different types of liquidity traders and trace their activity to different groups of informed traders.

Easley and O'Hara (1992) identify the spread as the key linking variable between informed and liquidity traders. In their theoretical model, market makers respond to increased trading by informed traders by setting the spread to have higher widths. The width gets adjusted downwards as market makers obtain more information about the identity of informed and liquidity trades. The width of the spread is a proxy for the extent of uncertainty the market maker feels about his profits. It may be advisable to enter the spread in the vector autoregressive model in lieu of prices, as it is a direct measure of the actions of the market maker in setting prices, rather than our more indirect measure of the price changes resulting from those spreads. The study may then be extended to dividends and earnings announcements as other information-based events in which informed trading may be forecasted by examining levels of uninformed trading.

REFERENCES

Admati, A. R., & Pfleiderer, P. (1988). A theory of intraday patterns: Volume and price variability. *Review of Financial Studies*, *1*, 3–40.

Bamber, L. S., Barron, O. E., & Stober, T. L. (1999). Differential interpretations and trading volume. *Journal of Financial and Quantitative Analysis*, *34*, 369–386.

Black, F. (1986). Noise. *Journal of Finance*, *41*, 529–543.

Cao, C., Chen, C., & Griffin, T. (2005). Informational content of option volume prior to takeover. *Journal of Business*, *78*, 1073–1109.

De Fontnouvelle, P., Fishe, P., & Harris, J. H. (2003). The behavior of bid-ask spreads and volume in options markets during the competition for listings in 1999. *Journal of Finance*, *58*, 2437–2464.

Easley, D. O., & O'Hara, M. (1992). Adverse selection and large trade volume: The implications for market efficiency. *Journal of Financial and Quantitative Analysis*, *27*, 185–208.

Glosten, L., & Milgrom, P. (1985). Bid, ask, and transaction prices in a specialist market with heterogeneously informed traders. *Journal of Financial Economics*, *14*, 71–100.

Heflin, F., & Shaw, K. W. (2005). Trade size and informed trading: Which trades are "big"? *Journal of Financial Research*, *28*, 133–163.

Kanodia, C., Bushman, R., & Dickhaut, J. (1986). Private information and rationality in the sunk costs phenomenon. Working Paper. University of Minnesota.

Kyle, A. S. (1985). Continuous auctions and insider trading. *Econometrica*, *53*, 1315–1335.

Lee, J., & Yi, C. H. (2001). Trade size and information-motivated trading in options and stock markets. *Journal of Financial and Quantitative Analysis*, *36*, 485–501.

Mitchell, M., Pulvino, T., & Stafford, E. (2004). Price pressure around mergers. *Journal of Finance*, *59*, 31–63.

Trueman, B. (1988). A theory of noise trading in securities markets. *Journal of Finance*, *43*, 83–95.

This page intentionally left blank

USING DATA ENVELOPMENT ANALYSIS (DEA) TO FORECAST BANK PERFORMANCE

Ronald K. Klimberg, Kenneth D. Lawrence and Tanya Lal

ABSTRACT

Forecasting is an important tool used by businesses to plan and evaluate their operations. One of the most commonly used techniques for forecasting is regression analysis. Often forecasts are produced for a set of comparable units which could be individuals, groups, departments, or companies that perform similar activities such as a set of banks, a group of mangers, and so on. We apply a methodology that includes a new variable, the comparable unit's data envelopment analysis relative efficiency, into the regression analysis. This chapter presents the results of applying this methodology to the performance of commercial banks.

INTRODUCTION

Quantitative forecasting models, even rather sophisticated models, are easier to develop and use today as a result of our improving computer technology. These quantitative forecasting techniques use historical data to predict the future. Most quantitative forecasting techniques can be categorized into either

Advances in Business and Management Forecasting, Volume 6, 53–61

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006004

time series approaches or causal models. Time series forecasting techniques are forecasting techniques that only use the time series data itself and not any other data to build the forecasting models. These time series approaches isolate and measure the impact of the trend, seasonal, and cyclical time series components. Causal models use a set of predictor/independent variables, possibly also including the time series components, which are believed to influence the forecasted variable. One of the most popular causal model approach is regression analysis. Regression techniques employ the statistical method of least squares to establish a statistical relationship between the forecasted variable and the set of predictor/independent variables.

Many forecasting situations involve producing forecasts for comparable units. A comparable unit could be an individual, group of individuals, a department, a company, and so on. Each comparable unit should be performing similar set of tasks. When applying regression analysis, the established statistical relationship is an average relationship using one set of weights assigned to the predictor/independent variables. However, when regression is applied to a set of comparable units the relative weight/importance of each of the predictor/independent variables will most likely vary from comparable unit to comparable unit. For example, if advertising is an independent variable, one comparable unit might emphasize advertising more (or less) than other comparable units. Either way is not necessarily better nor worse, it is just how that particular comparable unit emphasizes advertising. As a result, in some cases, the regression model could provide forecast estimates that are too high or too low.

In this chapter, we will apply and extend some of our recent previous work, Klimberg, Lawrence, and Lawrence (2005, 2008), in which we introduced a methodology that incorporates into the regression forecasting analysis a new variable that captures the unique weighting of each comparable unit. This new variable is the relative efficiency of each comparable unit that is generated by a nonparametric technique called data envelopment analysis (DEA). The next section provides a brief introduction to DEA. Subsequently, the methodology is discussed and the results of applying our methodology to a data set of commercial banks are presented. Finally, the conclusions and future extensions are discussed.

DATA ENVELOPMENT ANALYSIS (DEA)

DEA utilizes linear programming to produce measures of the relative efficiency of comparable units that employ multiple inputs and outputs.

DEA takes into account multiple inputs and outputs to produce a single aggregate measure of relative efficiency for each comparable unit. The technique can analyze these multiple inputs and outputs in their natural physical units without reducing or transforming them into some common measurement such as dollars.

The Charnes, Cooper, and Rhodes (CCR) DEA model (Charnes, Cooper, & Rhodes, 1978) is a linear program that compares the ratio of weighted outputs to weighed inputs, that is, efficiency, for each comparable unit. The efficiency of the kth comparable unit (i.e., E_k) is obtained by solving the following linear formulation:

$$\max E_k = \sum_{r=1}^{t} u_r Y_{rk}$$

s.t.

$$\sum_{i=1}^{m} v_i X_{ik} = 1$$

$$\sum_{r=1}^{t} u_r Y_{rj} - \sum_{i=1}^{m} -v_i X_{ij} \leq 0 \quad j = 1, \ldots, n$$

$$u_r, v_i \geq \varepsilon \quad \forall r, i$$

where following are the *parameters*: Y_{rj} is the amount of the rth output for the jth comparable unit, X_{ij} the amount of the ith input for the jth comparable unit, t the number of outputs, m the number of inputs, n the number of comparable units, and ε a small infinitesimal value; and following are the *decision variables*: u_r is the weight assigned to the rth output and v_i the weight assigned to the ith input.

The CCR DEA formulation determines objectively the set of weights, u_r and v_i, that maximizes the efficiency of the kth comparable unit, E_k. The constraints require the efficiency of each comparable unit, including the kth comparable unit, not to exceed 1, and the weights, u_r and v_i, must be positive. A similar DEA formulation must be solved for each comparable unit. A comparable unit is considered relatively inefficient (i.e., $E_k < 1$) if it is possible to increase its outputs without increasing inputs or decrease its inputs without decreasing outputs. A comparable unit identified as being efficient (i.e., $E_k = 1$) does not necessarily imply absolute efficiency. It is only relatively efficient as compared to the other comparable units that are being considered. These efficiency ratings allow decision makers to identify which comparable units are in need of improvement and to what degree.

Each efficiency score measures the relative efficiency of the comparable unit. These efficiency scores can be used to evaluate performance of the comparable units and provide benchmarks. Nevertheless, besides each efficiency score being composed of a different set of inputs and outputs values, each comparable unit's efficiency score includes a unique set of weights. The DEA process attempts to find objectively the set of weights that will maximize a comparable unit's efficiency. Therefore, the DEA model has selected the best possible set of weights for each comparable unit. The variation of these weights from one comparable unit to the other comparable unit allows each comparable unit to have its own unique freedom to emphasize the importance of each of these input and output variables in their own way. How well they do this is measured by the efficiency score.

Since the Charnes et al.'s 1978 paper, there have been thousands of theoretical contributions and practical applications in various fields using DEA. DEA has been applied to many diverse areas such as health care, military operations, criminal courts, university departments, banks, electric utilities mining operations, and manufacturing productivity (Klimberg & Kern, 1992; Seiford, 1996; Seiford & Thrall, 1990).

REGRESSION FORECASTING METHODOLOGY

Our regression forecasting methodology is designed to be applied to a historical data set of multiple input and output variables from a set of comparable units (Klimberg et al., 2005, 2008). The methodology is a three-step process. The first step selects a dependent variable and if necessary reduces the number of input and output variables. Basically, one output variable is identified to be the principal/critical variable that will be needed to be forecasted, for example, sales, production, or demand. If the number of input and output variables is relatively large, similar to the goal in multiple regression, we follow the principle of parsimony and try to build a model that includes the least number of variables, which sufficiently explains the dependent variable. In DEA, the combined total of inputs and outputs included in the model should be no more than half the number of comparable units being compared in the analysis (Boussofiane, Dyson, & Thanassoulis, 1991; Golany & Roll, 1989). Golany and Roll (1989) suggest employing both qualitative and quantitative (including stepwise regression) techniques to identify the significant set of input and output variables. The second step is to run the DEA for each comparable unit using the identified significant input and output variables. We use these efficiency scores as surrogate measures of

the unique emphasis of the variables and of performance. The last step uses the principal/critical output variable as the regression-dependent variable, all the significant input variables plus the DEA efficiency score as regression-independent variables, and run a multiple regression. This regression model with the DEA efficiency variable should be superior, that is, should have a significantly lower standard error of the mean and increase R^2, to the regression model without the DEA efficiency score variable.

EXAMPLE

Commercial banks are defined as "those banks whose business is derived primarily from commercial operations but which are also present in the retail banking and small and medium industry sectors" (Datamonitor, 2007). As of 2007, the commercial banking industry in the United States had a market value of $502.7 billion and was projected to grow by 23.7% to $622 billion in 2011 (Datamonitor, 2007). The different banks in the industry serve the same range of clients and offer the same services and therefore have limited competitive advantage on those fronts. Most companies derive competitive advantage from pricing and market reach. This has caused a trend toward consolidation in the industry. Since the balance of risk and return is crucial to profitability in commercial banking, consolidation has raised concerns about whether the large scale of some of the companies in the industry allows for proper oversight and regulation of risk.

Revenue is derived from two sources, interest income on deposits and noninterest income such as fees and commissions. Typically 50% or more of commercial banks' operating costs can be due to employee compensation. This is because of a need to have a large street branch presence and to competitively compensate higher level employees such as asset managers who are in high demand (Datamonitor, 2007).

Monitoring the performance of commercial banks is important to several parties. Customers are concerned about the safety of their deposits and access to affordable credit; shareholders are concerned about returns on their investment; managers are concerned with profitability; and finally regulators are concerned owing to the banks' role in the economy as the main source of financial intermediation and as custodians of a large portion of the nation's cash (Data Envelopment Analysis and Commercial Bank Performance: A Primer With Applications to Missouri Banks, 1992). Historically financial ratios such as return on assets (ROA) or return on investment (ROI) have been used to measure bank's performance. Although

financial ratios are useful for benchmarking purposes, there are multiple factors that contribute to a bank's performance at any given point in time (Seiford, 1996). To predict profitability through revenues or profits, it is crucial to understand the dynamics between the different resources used by banks and their relationship to profitability. Those resources include assets, equity, and number of employees.

Seiford and Zhu (1999) applied DEA to the 55 U.S. commercial banks that appeared in the *Fortune* 1000 list in April 1996. The DEA input variables were the number of employees, assets, and stockholder's equity; and the DEA output variables were revenue and profit. The selection of these variables were "based on *Fortune's* original choice of factors for performance characterization" (Seiford & Zhu, 1999).

We retrieved the same *Fortune* 1000 list of U.S. commercial banks from 2003 to 2007. We ran similar DEA models, that is, same input and output variables as Seiford and Zhu, for 2003–2006. Table 1 lists the frequency distribution of the DEA efficiency scores for these years. As shown in Table 1, these efficiency scores are rather dispersed.

Using the DEA efficiency scores as an input and revenue as our primary output variable, we ran regression models for 2004–2007. The basic regression equation used was

$$\text{revenue}(t) = \text{employees}(t-1) + \text{assets}(t-1) + \text{equity}(t-1) + \text{DEA}(t-1)$$

where $t = 2004, 2005, 2007$ (we refer to this model as w/DEA). Additionally, we ran the same regression without the DEA efficiency score variable (we refer to this model as NoDEA).

Table 1. Frequency Distribution of the DEA Efficiency Scores for Each Year.

Interval	Year			
	2003	2004	2005	2006
100	2	4	6	6
90.01 to 99.99	3	1	2	2
80.01 to 90	6	5	9	7
70.01 to 80	6	10	6	8
60.01 to 70	9	4	3	3
50.01 to 60	2	2	2	
40.01 to 50	1	2		
<40		1		

Table 2. The Regression R^2 Values for Each Year and for the Two Models.

R^2	Year			
	2004	2005	2006	2007
NoDEA	99.26	99.89	99.63	98.87
w/DEA	99.48	99.90	99.70	99.14
Difference	0.21	0.01	0.07	0.27

Table 3. The Regression Standard Errors for Each Year and for the Two Models.

Standard Error	Year			
	2004	2005	2006	2007
NoDEA	2,163.41	1,042.98	2,394.12	4,643.00
w/DEA	1,862.84	1,019.64	2,191.75	4,142.95
Decrease	300.57	23.34	202.37	500.05
% Decrease	13.89	2.24	8.45	10.77

Tables 2 and 3 summarize the regression models results with R^2 values and standard errors. The w/DEA models were consistently better than the NoDEA models. In terms of R^2 values, the NoDEA models had extremely high R^2 values every year. The w/DEA models only slightly increase the R^2 values; averaging only 0.14% improvement.

The standard error values for the w/DEA models, in Table 3, had a more significant improvement; averaging 8.84% decrease in the standard errors.

Table 4 summarizes the residual results by displaying the maximum and minimum residual for each model. In each case, the w/DEA regression models performed better than the NoDEA regression models.

CONCLUSIONS

In this chapter, we applied a new regression forecasting methodology to forecasting comparable units. This approach included in the regression analysis a surrogate measure of the unique weighting of the variables and of

Table 4. Residual Analysis for Each Year and for the Two Models.

	2004			2005		
	NoDEA	w/DEA	Improvement	NoDEA	w/DEA	Improvement
Maximum	6,346.03	3,756.31	2,589.72	1,930.83	1,699.98	565.30
Minimum	−5,469.49	−4,638.14	−1,906.04	−1,787.71	−1,720.04	−700.79
Average			102.38			6.28
	2006			2007		
Maximum	8,697.21	6,902.34	1,794.87	8,098.02	7,942.76	6,692.83
Minimum	−3,343.11	−3,085.70	−1,928.11	−13,852.49	−13,147.16	−2,539.06
Average			33.74			338.95

performance. This new variable is the relative efficiency of each comparable unit that is generated by DEA. The results of applying this new regression forecasting methodology including a DEA efficiency variable to a data set demonstrated that this may provide a promising rich approach to forecasting comparable units. We plan to perform further testing with other data sets, some with more comparable units and more years of data.

REFERENCES

Boussofiane, A., Dyson, R. G., & Thanassoulis, E. (1991). Applied data envelopment analysis. *European Journal of Operational Research*, *52*(1), 1–15.

Charnes, A., Cooper, W. W., & Rhodes, E. (1978). Measuring efficiency of decision making units. *European Journal of Operational Research*, *2*, 429–444.

Data Envelopment Analysis and Commercial Bank Performance: A Primer with Applications to Missouri Banks. (1992). Federal Reserve Bank of St. Louis Review, January, pp. 31–45.

Datamonitor. (2007). Commercial Banking in the United States, May. New York. Available at www.datamonitor.com

Golany, B., & Roll, Y. (1989). An application procedure for DEA. *Omega*, *17*(3), 237–250.

Klimberg, R. K., Lawrence, K. D., & Lawrence, S. M. (2005). Forecasting sales of comparable units with data envelopment analysis (DEA). *Advances in Business and Management Forecasting*, *4*, 201–214. JAI Press/North Holland.

Klimberg, R. K., Lawrence, K. D., & Lawrence, S. M. (2008). Improved performance evaluation of comparable units with data envelopment analysis (DEA). *Advances in Business and Management Forecasting*, *5*, 65–75. Elsevier Ltd.

Klimberg, R. K., & Kern, D. (1992). *Understanding data envelopment analysis (DEA)*. Working Paper no. 92-44. Boston University School of Management.

Seiford, L. M. (1996). Data envelopment analysis: The evaluation of the state of the art (1978–1995. *The Journal of Productivity Analysis*, *9*, 99–137.

Seiford, L. M., & Thrall, R. M. (1990). Recent developments in DEA: The mathematical programming approach to frontier analysis. *Journal of Econometric*, *46*, 7–38.

Seiford, L. M., & Zhu, J. (1999). Profitability and marketability of the top 55 U.S. commercial banks. *Management Science*, *45*(9), 1270–1288.

This page intentionally left blank

PART II
MARKETING AND DEMAND APPLICATIONS

This page intentionally left blank

FORECASTING DEMAND USING PARTIALLY ACCUMULATED DATA

Joanne S. Utley and J. Gaylord May

ABSTRACT

This chapter uses advance order data from an actual manufacturing shop to develop and test a forecast model for total demand. The proposed model made direct use of historical time series data for total demand and time series data for advance orders. Comparison of the proposed model to commonly used approaches showed that the proposed model exhibited greater forecast accuracy.

INTRODUCTION

In many businesses, a portion of the demand for a future time period may be known due to advance customer orders. For example, at any point in time, hotel reservations provide partial information about the customer demand that will actually be realized in future time periods. Similarly, in a manufacturing shop, firm orders to date with customer designated lead times allow the manufacturer to know with certainty a portion of the actual demand in future time periods.

Advances in Business and Management Forecasting, Volume 6, 65–80

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006005

Past research has shown that if a forecaster can utilize this kind of partial demand information in developing a forecast, he has the opportunity to produce a more accurate forecast than if he relied on historical data alone (Kekre, Morton, & Smunt, 1990). Past research has also suggested that the forecast methodology should not be so complex that it would be difficult to implement in practice (Guerrero & Elizondo, 1997). This chapter proposes a forecast model, which is straightforward in approach and relatively easy to implement. The proposed model utilizes both partial demand data and historical time series data to forecast customer demand.

This chapter is organized as follows. The following section provides an overview of forecast methods which use advance customer order data. The "Case Study" section describes the forecast model devised by the authors. It illustrates the application of the proposed model at an actual manufacturing company and develops alternative forecasts through commonly used approaches. The accuracy of the proposed forecast model is then compared to the accuracy of the standard models. The chapter concludes with a discussion of the research findings and suggestions for future research.

OVERVIEW OF THE LITERATURE

One of the earliest methods of forecasting with partial data was proposed by Murray and Silver (1966) to address the style goods problem. In this problem, there is a finite selling period for the item, sales vary in a seasonal and somewhat predictable pattern, and the item can be bought or produced only on a limited number of occasions. Murray and Silver (1966) noted that although there is initially great uncertainty about the sales potential of the item, the manager can improve the forecast as sales become known over time. Murray and Silver (1966) utilized a dynamic programming model in which each state represented both the amount of unsold inventory and current knowledge of sales. Green and Harrison (1973) also used Bayesian methods to solve the style goods problem. Their approach was utilized in a mail order company to forecast sales for a sample of 93 different types of dresses. Guerrero and Elizondo (1997) have noted that these early studies adopted a Bayesian perspective to the problem because information on future demand was extremely limited. Guerrero and Elizondo (1997, p. 879) also observed that a major difficulty with this approach is that "the resulting methods usually require a large amount of expertise from the analyst to implement them."

Bestwick's (1975) study was the first in the forecasting literature to investigate a simpler alternative to Bayesian type models. Bestwick's (1975) approach utilized a multiplicative model, which assumed that total demand for a future time period t can be found at an earlier time period $t-h$ by dividing the advance orders for period t, booked h or more periods ahead, by a cumulative proportion C_h. The cumulative proportion C_h can be interpreted as the percent of total actual demand for a time period realized through accumulated customer orders, h periods in advance. For example, if a manager knows that at time $t-h$, 25% of the total demand for period t will already be known, and if the sum of the advance orders for period $t = 100$ units, then the forecast for period t is $100/.25 = 400$ units.

Bestwick (1975) used his multiplicative model in conjunction with a forecast monitoring procedure, which included traditional elements of statistical process control such as cusum techniques, the mean chart, and the sampling distribution of the range. Bestwick (1975) reported that his methodology was successfully implemented in a number of areas including demand forecasting, inventory control, budgeting, production control, and turnover analysis. Today, the multiplicative model remains popular in both the private and public sector (Kekre et al., 1990). Its simplicity continues to make it attractive to managers who lack the technical expertise or financial resources needed to implement more sophisticated forecast models.

Despite its popularity, Bestwick's (1975) multiplicative model does exhibit some shortcomings. First, this approach assumes that each cumulative proportion C_h will remain constant over time; however, drift in the C_h values can often occur in practice. Second, this model does not use the total demand time series directly in the forecast. Thus, the simple multiplicative model is not particularly responsive to changes in total demand (Bodily & Freeland, 1988). Finally, the accuracy of the multiplicative method is not only a function of the stability of the C_h values but also their accuracy. The accuracy of the C_h values tends to decline as h increases. When $h > L$, the maximum customer specified lead time, the forecaster cannot continue to use this method since no customer order data will exist this far in advance (Kekre et al., 1990).

Kekre et al. (1990) addressed the problem of variability of the cumulative proportions in Bestwick's model by using exponential smoothing to update the C_h values. They used partially known demand data from a printing firm to test their model, assuming a 5-day forecast horizon. Kekre et al. (1990) found that the multiplicative model outperformed a simple exponential smoothing model that used time series data for total demand. They also examined an additive model for partial demand data. The additive model

assumed that a forecast for total demand could be found by adding the known portion of demand for a future time period with the smoothed unknown portion of demand for the future time period. Unlike the smoothed multiplicative model, the additive model assumes that the known portion of demand contains no information about the unknown portion of demand (Kekre et al., 1990).

Kekre et al. (1990) found that both the smoothed multiplicative and additive models performed better than traditional exponential smoothing for a planning horizon of four periods or less, even if random demand shocks occurred. Kekre et al. (1990, p. 123) also reported that the smoothed additive model was more appropriate for long lead times than the multiplicative approach and observed that the smoothed additive model "becomes indistinguishable from exponential smoothing as the lead time increases." They concluded their paper by suggesting that future research could focus on the combination of multiplicative and additive techniques through regression analysis.

Bodily and Freeland (1988) argued that alternatives to the simple and smoothed multiplicative models were needed, especially since both models fail to use the total demand time series directly. Using simulated booking and shipment data, Bodily and Freeland (1988) tested six partial demand forecast models. The first two models were the simple multiplicative model and the smoothed multiplicative model discussed earlier. The third model combined smoothed shipments with fixed C_h factors, whereas the fourth model combined smoothed shipments with smoothed C_h factors. The fifth and sixth models used Bayesian shipments with fixed C_h factors and smoothed C_h factors, respectively. Results of the model comparisons showed that the smoothed multiplicative model performed the best overall.

In a later paper, Guerrero and Elizondo (1997) used least squares estimation to forecast total demand with partially accumulated data. Their model specified a set of L linear regressions, where L is the longest lead time of the forecast. Guerrero and Elizondo (1997) used both Kekre et al.'s (1990) data set and data on the Mexican economy to compare the accuracy of their approach to that of algorithmic solutions. They reported that their statistical approach was more accurate for all lead times.

A recent paper by Waage (2006) used a stochastic dynamic model to forecast demand with advance order data. The model combined two information sources (1) an econometric sales forecast and (2) partially known information about customer orders for future time periods. Waage (2006) applied his model to the problem of forecasting sales for special purpose computers. Waage (2006, p. 24) reported that the sales forecast

produced by his model converged on the actual sales trajectory even before the actual trajectory was known.

The forecast models summarized in this section reflect varying degrees of complexity and forecast accuracy. Although some companies may prefer to implement more sophisticated forecast models when partial order data are available, other business, particularly smaller businesses, may prefer approaches that are simpler and require less expertise. The next section of this chapter will describe a forecast approach that is relatively simple to use yet exploits the availability of both advance order data and time series data for total demand.

CASE STUDY

An electronics component manufacturer located in the southeast United States provided the research context for this study. Although the company produces a variety of products, order data for a single product, referred to as product (A), will be used to illustrate the forecast model. Analysis of 9 months of historical data for this product showed that each customer order included a requested delivery date (or customer designated lead time) in addition to the order quantity. Designated lead times typically ranged from 1 to 4 months, although a lead time of 5 or 6 months would occasionally be requested. A customer's order quantity and designated lead time did not remain constant over time; instead, they varied with each order.

The manufacturer wished to forecast total demand for a 6 month planning horizon (months 10–15) by using the partial order data it already possessed at the end of month 9. The authors developed the model shown in Exhibit 1 to make use of both the partial order data available for months 1–13 and the total demand data for months 1–9 in preparing the forecasts. The remainder of this section will use these data to illustrate concepts contained in the model's design.

For a particular product, we will consider all customer orders that are requested for delivery in a specific time period (t). This demand may be distributed with respect to the lead times supplied by the customer. Table 1 shows such a distribution of orders for product (A), which was actually received by the manufacturing company.

Let t be a particular time period and h be a designated number of such periods. $D(t,h)$ shall denote the sum of customer demands for period (t), which have lead times $\geq h$. As an example, in Table 1, let t correspond to month (4) of 2005. For this request date, orders were received from month

Exhibit 1. Ratio Model for Total Demand Forecasts.

Let	
t	a particular time period
h	a designated number of time periods
$D(t,h)$	accumulated demand for period t occurring h or more periods in advance of t (or the sum of orders for period t in which the customer supplied lead time $\geq h$
$D(t)$	the total demand for period t
$F(t)$	the forecast for total demand for period t
$R(t) = D(t)/D(t-1)$	the ratio of total demand for period t to total demand for period $t-1$
$FR(t)$	the forecast ratio of total demand in period t to total demand in period $t-1$
$P(t,h) = D(t,h)/D(t)$	the ratio of partially accumulated for period t, known h or more periods in advance, to total demand for period t
$FP(t,h)$	the forecast of the ratio of partially accumulated demand for period t, known h or more periods in advance, to total demand for period t

For each t in the forecast horizon, the forecast for $R(t)$ is given by
$FR(t) = (D(t,h))/(D(t-1,h)) \bullet (FP(t-1,h))/(FP(t,h))$, where h is smallest lead time for which $FP(t,h)$ can be computed

The forecast for total demand for period t is given by
$F(t) = FR(t)F(t-1)$

(10) of 2004 to month (5) of 2005. These orders provided lead times from +6 to −1. $D(t,-1) = 266$ is the total demand that ultimately materialized. $D(t,4) = 106$ was realized with orders received through month (12) of 2004. (The existence of a negative lead time illustrates that customer demand records were not always accurately maintained.)

Each column in Table 1 shows customer demand as it is distributed with respect to lead time. Characteristically, these distributions range over lead times from approximately 6 to −1 with a maximum demand occurring at 3 or 4. Let $\bar{t}$ denote the current time period (that period from which we shall project our forecast). In our example $\bar{t} = 9$, which corresponds to month (11) of 2005. Customer demand $D(\bar{t})$ is completely known. $D(\bar{t}+1)$ identifies demand for the first period in our forecast horizon. We wish to forecast for each t over a horizon of six time periods. The model was developed by using the following observations.

Observation I: Often, customer demand (sales) of a particular product within a business will grow or decline each time period as a percentage of previous values rather than by a fixed amount. Should this occur, $D(t) = k \bullet D(t-1)$, where $k \bullet 100$ is the percentage value. The proposed model does not require k to remain fixed but does retain the expectation that demand ratios,

Table 1. Distribution of Product Demand with Respect to Lead Time.

Product (A)		Time Period (t)	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
		Date Requested	2005										2006				
			3	4	5	6	7	8	9	10	11	12	1	2	3	4	5
2004	10		12	1													
	11		139	8													
	12		107	97	17	1											
Data received 2005	1		48	115	165	22	13										
	2		7	37	59	105	29		2								
	3		1	2	9	40	157	35			2						
	4		3	5	6	21	38	143	74	5	4						
	5		1	1	4	4	3	7	75	32							
	6		3					1	5	53	29						
	7						1	2	11	20	18	1					
	8						6		3	46	115	70	32	1			
	9						4	4	4	8	38	43	32				
	10						2	1	8	22	31	54	132	17	1		1
Current date	11									24	25	22	97	45	8		
Customer demand $D(t,h)$			321	266	260	193	253	193	182	210	262	190	293	63	9		1
Current minimum lead time h			−3	−1	0	−1	−3	−2	−1	−1	0	1	2	3	4	5	6

$R(t) = D(t)/D(t-1)$, are linearly related over time. Fig. 1 shows a plot of total demand, $D(t)$, for the product (A) data. This plot does not appear linear.

Fig. 2 shows a plot of the corresponding demand ratios, which does display a linear characteristic. The model forecasts demand ratios over the

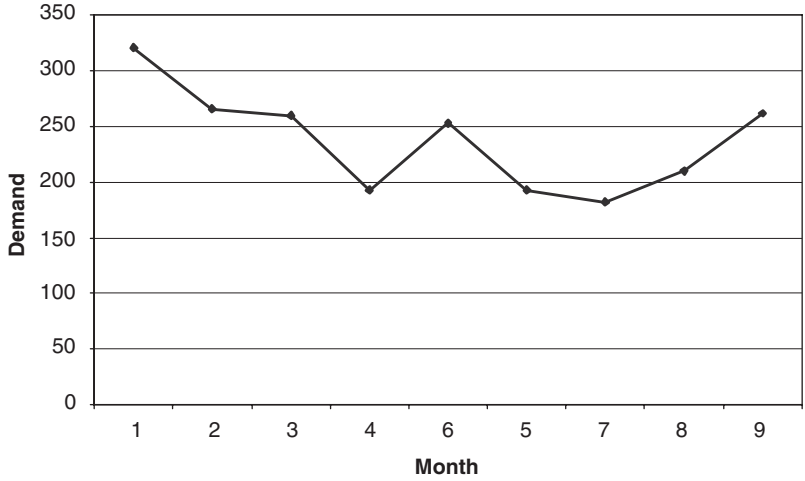

Fig. 1. $D(t)$ = Total Demand.

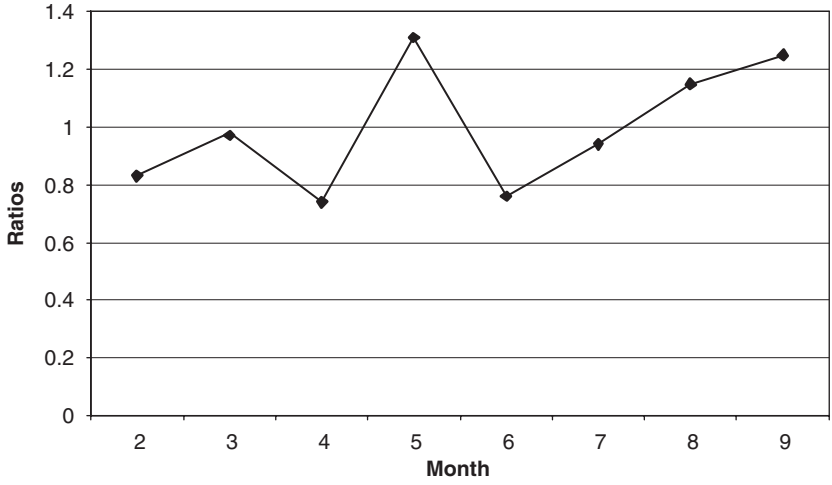

Fig. 2. $D(t)/D(t-1)$ = Demand Ratios.

six period horizon. A final calculation converts these ratios into a forecast of demand. This conversion uses the recursive formula: $F(t) = FR(t) \bullet F(t-1)$.

Observation II: As explained in the "Overview of the Literature" section, the basic multiplicative model customarily assumes that $D(t,h)/D(t)$, the percent of total demand for period t with a lead time $\geq h$, remains constant over time. Examination of product (A) data showed that, for fixed h, $P(t,h) = D(t,h)/D(t)$ *does not* remain stationary. As shown in Table 2, the percent of total demand declines as the time periods advance.

To forecast $D(t)/D(t-1)$ over time periods 10, 11, 12, 13 where lead time demand exists, we must first forecast $P(t,h)$ for $h = 1, 2, 3, 4$. These forecasts were obtained using exponential smoothing. They are shown in Table 2 in rows 10–13. Let $FP(t,h)$ denote these forecast values. We estimate $D(t,h)/D(t) = FP(t,h)$ or $D(t) = D(t,h)/FP(t,h)$. For each t, we use the smallest h for which $FP(t,h)$ can be computed. We have

$$\frac{D(t)}{D(t-1)} \approx \frac{D(t,h)}{D(t-1,h)} \bullet \frac{FP(t-1,h)}{FP(t,h)} = FR(t), \quad \text{where } t = 10,\ 11,\ 12,\ 13$$

These forecasts of demand ratios are shown in the right most column of Table 2. Fig. 3 shows a plot of these four lead time ratios, which appear as a continuation of the ratios shown in Fig. 2.

Observation III: Time periods 14 and 15 are in the forecast horizon but do not have partially accumulated demand data. To obtain ratio forecasts for these two periods, exponential smoothing was applied to the existing ratios. These smoothed ratios are plotted in Fig. 4 and their values are displayed in Table 3.

The smoothed ratios for periods 14 and 15 together with the lead time ratios for periods 10, 11, 12 and 13 form the ratio forecasts over the six period horizon. All of these ratios are shown in Table 3.

Table 4 shows the results of converting the complete ratio forecasts into total demand forecasts. Also shown is the complete demand that actually materialized after the 6-month horizon had transpired.

Fig. 5 plots a comparison of the actual demand with the forecast of demand obtained from converting the ratio forecasts. Over the 4-month lead time, a "seasonal pattern" existed for product (A) due to budget considerations of the customers. There was a decline in demand for December (period 10) followed by a sharp increase in January and a subsequent decline in February and March. This pattern was anticipated in November (current time period 9) using lead time ratios.

Table 2. Calculation of Lead Time (LT) Ratios.

Month	Total Demand ($D(t)$)	Demand ($D(t,1)$ LT> = 1)	% of Total ($P(t,1)$ LT> = 1)	Demand ($D(t,2)$ LT> = 2)	% of Total ($P(t,2)$ LT> = 2)	Demand ($D(t,3)$ LT> = 3)	% of Total ($P(t,3)$ LT> = 3)	Demand ($D(t,4)$ LT> = 4)	% of Total ($P(t,4)$ LT> = 4)	Demand Ratios ($D(t)/D(t-1)$)
1	321	313	0.975	306	0.953	258	0.804	151	0.470	
2	266	260	0.977	258	0.970	221	0.831	106	0.398	0.83
3	260	256	0.985	250	0.962	241	0.927	182	0.700	0.98
4	193	193	1.000	189	0.979	168	0.870	128	0.663	0.74
5	253	240	0.949	240	0.949	237	0.937	199	0.787	1.31
6	193	188	0.974	186	0.964	185	0.959	178	0.922	0.76
7	182	170	0.934	167	0.918	156	0.857	151	0.830	0.94
8	210	164	0.781	156	0.743	110	0.524	90	0.429	1.15
9	262	237	0.905	206	0.786	168	0.641	53	0.202	1.25
10		190	0.841	168	0.753	114	0.542	71	0.343	0.86
11				293	0.709	196	0.452	64	0.257	1.85
12						63	0.361	18	0.171	0.4
13								9	0.085	1.01

$$\frac{D(t,h)}{D(t)} = P(t,h) \qquad \frac{D(t)}{D(t-1)} = \frac{D(t,h)}{D(t-1,h)} \bullet \frac{P(t-1,h)}{P(t,h)}$$

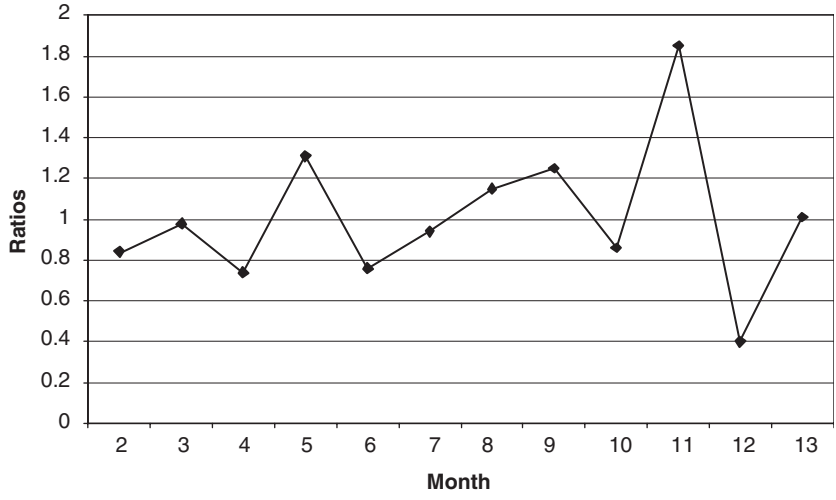

Fig. 3. Actual and Lead Time Ratios.

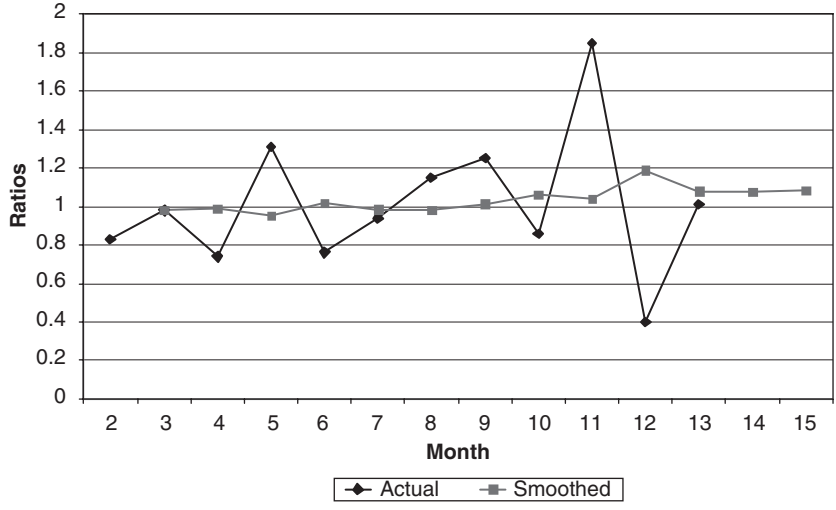

Fig. 4. Smoothed Ratio Forecasts.

The expected demand forecast generated from lead time information may be compared with forecasts of expected demand when no ratio or lead time information was used. Table 5 and Fig. 6 show the results when forecasts used only complete demand up to the current time period 9.

Table 3. Complete Forecast of Demand Ratios.

Month	Demand Ratios	Smoothed Ratio Forecast $\alpha = .15$, $\beta = .05$	Complete Forecast of Demand Ratios
2	0.83		0.83
3	0.98	0.98	0.98
4	0.74	0.99	0.74
5	1.31	0.95	1.31
6	0.76	1.02	0.76
7	0.94	0.98	0.94
8	1.15	0.98	1.15
9	1.25	1.01	1.25
10	0.86	1.06	0.86
11	1.85	1.04	1.85
12	0.4	1.19	0.4
13	1.01	1.08	1.01
14		1.08	1.08
15		1.08	1.08

Table 4. Total Demand Forecast.

Month	Total Demand	Complete Ratio Forecasts	Total Demand Forecasts	Actual Total Demand
1	321		321	321
2	266	0.83	266	266
3	260	0.98	260	260
4	193	0.74	193	193
5	253	1.31	253	253
6	193	0.76	193	193
7	182	0.94	182	182
8	210	1.15	210	210
9	262	1.25	262	262
10		0.86	225	235
11		1.85	416	405
12		0.40	166	264
13		1.01	168	206
14		1.08	181	212
15		1.08	195	232

A two parameter exponential model was used. Three forecasts are shown using three sets of values for α and β. The forecast that generated the smallest mean squared error was the one which produced the lowest expected demands. Without the use of lead time information there was no

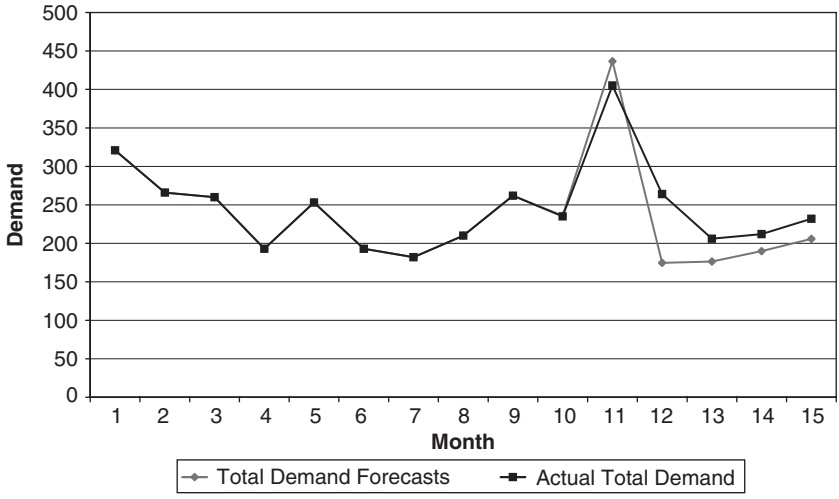

Fig. 5. Demand: Actual vs. Ratio Forecasts of Total Demand.

Table 5. Exponentially Smoothed Forecasts for Total Demand.

Observations	Actual Demand	Forecast-1 $\alpha = .15$; $\beta = .05$	Forecast-2 $\alpha = .4$; $\beta = .3$	Forecast-3 $\alpha = .5$; $\beta = .6$
1	321	275.11111	275.1111	275.1111
2	266	273.00528	289.64	302.4889
3	260	262.91278	273.5206	277.7311
4	193	253.41234	259.8264	257.0329
5	253	234.83385	216.7908	193.9739
6	193	228.17838	219.3145	210.1522
7	182	213.25739	193.671	183.0957
8	210	198.69012	172.4843	163.7388
9	262	190.59277	175.4743	181.9387
10	235	192.04558	208.4513	241.057
11	405	182.78729	206.818	260.1447
12	264	173.52901	205.1848	279.2324
13	206	164.27072	203.5515	298.32
14	212	155.01244	201.9182	317.4077
15	232	145.75415	200.2849	336.4954
MSE		1,509.4	2,038.6	2,242.4

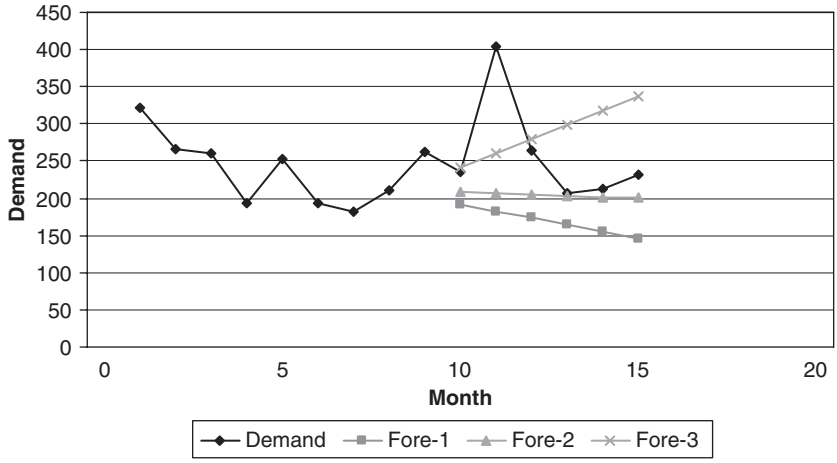

Fig. 6. Actual Demand vs. Exponentially Smoothed Forecasts.

Table 6. Comparison of Accuracy Measures: Ratio Model vs. Smoothed Multiplicative Model.

Month	Actual Demand	Ratio Forecast	Error Terms Ratio Model	Smoothed Multiplicative Forecast	Error Terms Multiplicative Model
10	235	225	10	210	25
11	405	416	−11	373	32
12	264	166	98	98	166
13	206	168	38	45	168
MAD			39.25		96
MSE			2,817.25		13,781.5
MAPE			15.7%		39.9%

anticipation that expected demand would sharply increase and then quickly decline within the forecast horizon.

In addition to comparing the proposed ratio model to exponential smoothing models, which did not incorporate any partial demand data, the authors also computed forecasts for periods 10–13 by using the smoothed multiplicative model. (It was not possible to develop forecasts for periods 14 and 15 with the multiplicative approach since sufficient partial demand data were not available for these time periods.) Table 6 shows that the ratio

model outperformed the smoothed multiplicative model on all three accuracy measures: the mean absolute deviation (MAD; 39.25 vs. 96, respectively), the mean squared error (MSE; 2817.25 vs. 13781.5, respectively), and the mean absolute percent error (MAPE; 15.7% and 39.9%, respectively).

DISCUSSION

Results from this study indicated that despite the very limited partial order data available for the second half of the forecast horizon, the ratio method outperformed the exponential smoothing models, which utilized only total demand time series data. Results also established the superiority of the ratio method over the smoothed multiplicative approach in this research context. There are several reasons why the ratio method outperformed the other approaches in this research context. First, the use of total demand ratios in computing the forecasts produced a smoothing effect on the data, thereby leading to greater forecast accuracy. Second, unlike the multiplicative approach, the ratio method made direct use of both the historical values and forecasted values of the total demand time series. Third, in contrast to the exponential smoothing models, the ratio model exploited the availability of the partially known order data.

The results presented in this chapter serve only to illustrate the proposed forecast model. Larger data sets and additional research contexts are needed to better study the effectiveness of the model. In addition, alternative forecast methodologies that are more complex than the smoothed multiplicative model or the exponentially smoothed model could be used for comparison purposes.

This study developed a forecast model for only one product made by the manufacturer. Also, the customer-specified lead times were specified in months. However, the results obtained suggest that the model could be applied to other products and services and that the customer-designated lead times period could be much shorter – perhaps weeks or even days.

REFERENCES

Bestwick, P. (1975). A forecast monitoring and revision system for top management. *Operational Research Quarterly*, *26*, 419–429.

Bodily, S., & Freeland, J. (1988). A simulation of techniques for forecasting shipments using firm order-to-date. *Journal of the Operational Research Society*, *39*(9), 833–846.

Green, M., & Harrison, P. (1973). Fashion forecasting for a mail order company using a Bayesian approach. *Operational Research Quarterly*, *24*, 193–205.

Guerrero, V., & Elizondo, J. (1997). Forecasting a cumulative variable using its partially accumulated data. *Management Science*, *43*(6), 879–889.

Kekre, S., Morton, T., & Smunt, T. (1990). Forecasting using partially known demands. *International Journal of Forecasting*, *6*, 115–125.

Murray, G., & Silver, E. (1966). A Bayesian analysis of the style goods problem. *Management Science*, *12*(11), 785–797.

Waage, F. (2006). Extracting forecasts from advance orders. *Advances in Business and Management Forecasting*, *4*, 13–26.

FORECASTING NEW ADOPTIONS: A COMPARATIVE EVALUATION OF THREE TECHNIQUES OF PARAMETER ESTIMATION

Kenneth D. Lawrence, Dinesh R. Pai and Sheila M. Lawrence

ABSTRACT

Forecasting sales for an innovation before the product's introduction is a necessary but difficult task. Forecasting is a crucial analytic tool when assessing the business case for internal or external investments in new technologies. For early stage investments or internal business cases for new products, it is essential to have some understanding of the likely diffusion of the technology. Diffusion of innovation models are important tools for effectively assessing the merits of investing in technologies that are new or novel and do not have prima facie, predictable patterns of user uptake. Most new product forecasting models require the estimation of parameters for use in the models. In this chapter, we evaluate three techniques to determine the parameters of the Bass diffusion model by using an example of a new movie.

Advances in Business and Management Forecasting, Volume 6, 81–91

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006006

INTRODUCTION

Forecasting new adoptions after a product introduction is an important marketing problem. The Bass model for forecasting is most appropriate for forecasting sales of an innovation (more generally, a new product) where no closely competing alternatives exist in the marketplace. Managers need such forecasts for new technologies or major product innovations before investing significant resources in them.

The Bass model offers a good starting point for forecasting the long-term sales pattern of new technologies and new durable products under two types of conditions (Lawrence & Geurts, 1984; Lilien & Rangaswamy, 2006).

(1) The firm has recently introduced the product or technology and has observed its sales for a few time periods; or
(2) The firm has not yet introduced the product or technology, but it is similar in some way to existing products or technologies whose sales history is known.

In this chapter, we introduce a forecasting model developed by Frank Bass that has proven to be particularly effective in forecasting the adoption of innovative and new technologies in the market place. We then use three techniques: nonlinear programming (NLP), linear regression (LR), and minimizing the sum of deviations to estimate the parameters of the Bass model and compare their performance using dataset of a new movie (Bass, 1969; Bass, 1993).

LITERATURE REVIEW

Rogers (1962) discussed the theory of adoption and diffusion of new products at length. The innovation adoption curve of Rogers classifies adopters of innovations into various categories, based on the idea that certain individuals are inevitably more open to adaptation than others. The various categories of adopters specified in the literature are (1) innovators, (2) early adopters, (3) early majority, (4) late majority, and (5) laggards (Fig. 1). The theory suggests that *innovators* are individuals who adopt new products independently of the decisions of other individuals in a social system. The literature aggregates groups (2) through (5) aforementioned and defines them as *imitators* (Bass, 1969). Imitators are adopters who are influenced in the timing of adoption by various external factors.

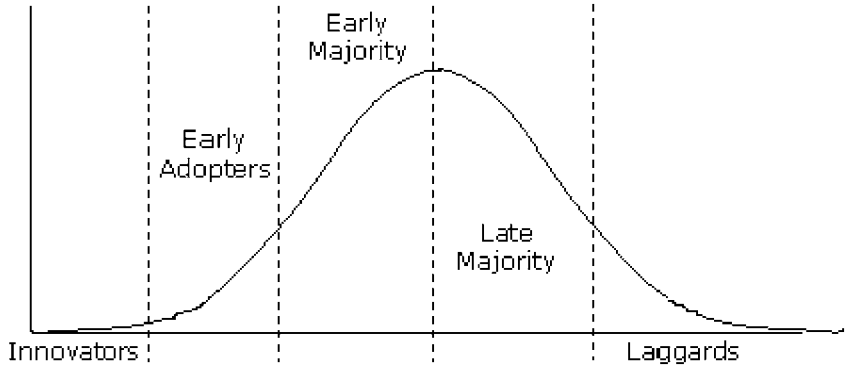

Fig. 1. Rogers Adoption/Innovation Curve.

The main impetus underlying diffusion research in marketing is the Bass model. The Bass model synthesizes the Fourt and Woodlock (1960) and Mansfield (1961) and employs a generalized logistic curve with these two models as its special cases. The Bass model assumes that potential adopters of an innovation are influenced by two means of communication – mass media and word of mouth. It further assumes that the adopters of an innovation comprise two groups: a group influenced by the mass media (external influence) and the other group influenced by the word-of-mouth communication (internal influence). Bass termed the first group "Innovators" and the second group "Imitators."

Since the publication of the Bass's new product growth model, research on the modeling of the diffusion of innovations in marketing has resulted in an extensive literature (Mahajan & Muller, 1979). The Bass model has been used for forecasting innovation diffusion in retail service, industrial technology, agricultural, educational, pharmaceutical, and consumer durable goods markets (Akinola, 1986; Bass, 1969; Dodds, 1973; Kalish & Lilien, 1986; Lancaster & Wright, 1983; Lawton & Lawton, 1979; Nevers, 1972; Tigert & Farivar, 1981).

THE BASS MODEL

The Bass model derives from a hazard function (the probability that an adoption will occur at time t given that it has not yet occurred). Thus, $f(t)/[1 - F(t)] = p + qF(t)$ is the basic premise underlying the Bass model.

Suppose that the (cumulative) probability that someone in the target segment will adopt the innovation by time t is given by a nondecreasing continuous function $F(t)$, where $F(t)$ approaches 1 (certain adoption) as t gets large. Such a function is depicted in Fig. 2(a), and it suggests that an individual in the target segment will eventually adopt the innovation. The derivative of $F(t)$ is the probability density function, $f(t)$ (Fig. 2(b), which indicates the rate at which the probability of adoption is changing at time t. To estimate the unknown function $F(t)$, we specify the conditional likelihood $L(t)$ that a customer will adopt the innovation at exactly time t since introduction, given that the customer has not adopted before that time (Lilien & Rangaswamy, 2006).

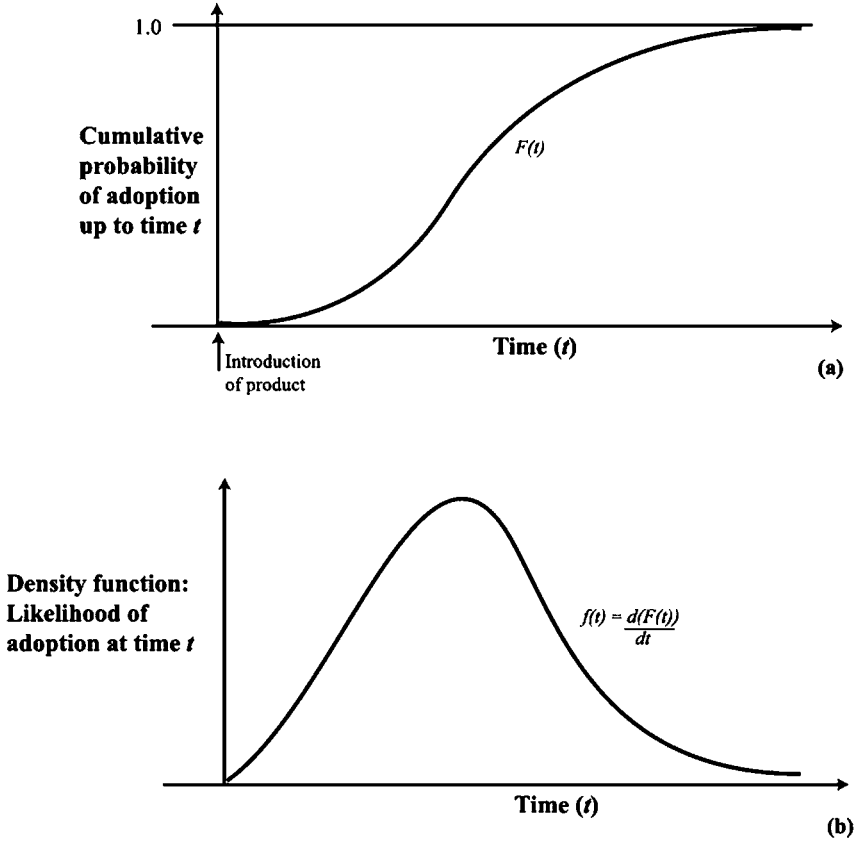

Fig. 2. Graphical Representation of the Probability of a Customer's Adoption of a New Product Over Time.

The conditional likelihood that a customer will adopt the innovation exactly at time t since introduction, given that the customer has not adopted before that time is (Lawrence & Lawton, 1981)

$$L(t) = \frac{f(t)}{1 - F(t)} \tag{1}$$

where $F(t)$ is a nondecreasing function, probability that someone in the target segment will adopt the innovation by time t; and $f(t)$ the rate at which the probability of adoption is changing at time t.

Bass (1969) proposed that $L(t)$ be defined to be equal to

$$L(t) = p + \frac{q}{m} C(t) \tag{2}$$

where $C(t)$ is the number of customers (or a multiple of that number, such as sales) who have already adopted the innovation by time t.

Following are the three parameters of the model that must be estimated:

m – A parameter representing the total number of customers in the adopting target segment, all of whom will eventually adopt the product. A company introducing a new product is obviously interested in the value of this parameter.
q – Coefficient of imitation (or coefficient of internal influence). This parameter measures the likelihood of adoption due to a potential adopter being influenced by someone who has already adopted the product. It measures the "word-of-mouth" effect influencing purchases.
p – Coefficient of innovation (or coefficient of external influence). This parameter measures the likelihood of adoption, assuming no influence from someone who has already purchased (adopted) the product. It is the likelihood of someone adopting the product due to her or his own interest in the innovation.

Let C_{t-1} be the number of people (or a multiple of that number, such as sales) who have adopted the product through time $t-1$. Therefore, $m-C_{t-1}$ is the number of potential adopters remaining at time $t-1$. We refer to time interval between time $t-1$ and time t as time period t.

The likelihood of adoption due to imitation is

$$q\left(\frac{C_{t-1}}{m}\right)$$

where C_{t-1}/m is the fraction of the number of people estimated to adopt the product by time $t-1$.

The likelihood of adoption due to innovation is simply p, the coefficient of innovation. Thus, the likelihood of adoption is

$$p + q\left(\frac{C_{t-1}}{m}\right)$$

Thus, FR_t, the forecast of the number of new adopters during time period t, is

$$FR_t = \left(p + q\left(\frac{C_{t-1}}{m}\right)\right)(m - C_{t-1}) \tag{3}$$

Eq. (3) is known as the Bass forecasting model.

Let S_t denote the actual sales in period t for $t = 1,2, \ldots, N$. The forecast in each period and the corresponding forecast error E_t is defined by

$$FR_t = \left(p + q\left(\frac{C_{t-1}}{m}\right)\right)(m - C_{t-1})$$

$$E_t = FR_t - S_t$$

THE THREE METHODS

Nonlinear Programming

NLP is used to estimate the parameters of the Bass forecasting model (Anderson, Sweeney, Williams, & Martin, 2005).

Minimizing the sum of errors squared, NLP formulation is

$$\min \sum_{t=1}^{N} E_t^2$$

$$\text{s.t.} \quad FR_t = \left(p + q\left(\tfrac{C_{t-1}}{m}\right)\right)(m - C_{t-1}) \quad t = 1, 2, \ldots, N \tag{4}$$

$$E_t = FR_t - S_t \quad t = 1, 2, \ldots, N \tag{5}$$

Linear Regression

Substituting Eq. (2) in Eq. (1), we get

$$f(t) = \left[p + \frac{q}{m}C(t)\right][1 - F(t)] \tag{6}$$

Simplifying Eq. (6), we get

$$S(t) = pm + (q - p)C(t) - \frac{q}{N}[C(t)]^2 \tag{7}$$

$$S(t) = a + bm(t - 1) + cm^2(t - 1) \tag{8}$$

We can then estimate the parameters (a, b, and c) of the linear function in Eq. (8) using OLS regression (Lilien & Rangaswamy, 2006).

We can calculate the Bass model parameters as

$$m = \frac{-b - \sqrt{b^2 - 4ac}}{2c}$$
$$p = \frac{a}{m}$$
$$q = p + b$$

We need the sales data for at least three periods to estimate the model. To be consistent with the model, $m > 0$, $b \geq 0$, and $c < 0$

Minimize the Least Absolute Deviation (LAD)

Here, we use the least absolute deviation (LAD) method to estimate the Bass model parameters. The method of LAD finds applications in many areas, due to its robustness over the least squares method. LADs are robust in that they are resistant to outliers in the data. The formulation for LAD through linear programming (Vanderbei, 2007)

$$\min \sum_{t=1}^{N} |FR_t - S_t| \tag{9}$$

Equivalent linear program:

$$\min \sum_{t=1}^{N} d_t \tag{10}$$

s.t.

$$FR_t = \left(p + q\left(\frac{C_{t-1}}{m}\right)\right)(m - C_{t-1}) \quad t = 1, 2, \ldots, N \tag{11}$$

$$d_t + FR_t - S_t \geq 0 \quad t = 1, 2, ..., N \tag{12}$$

$$d_t - FR_t + S_t \leq 0 \quad t = 1, 2, ..., N \tag{13}$$

where $d_t \geq 0$ is the deviation or the forecast error.

THE EXAMPLE

We consider an example of the box office revenues (in $ millions) of a movie over the first 12 weeks after release (Lilien & Rangaswamy, 2006). Fig. 3 shows the graph of the revenues for the movie. Though, the box office revenues for time period t are not the same as the number of adopters during time period t, the revenues are a multiple of the number of moviegoers as the number of repeat customers are low. From the graph,

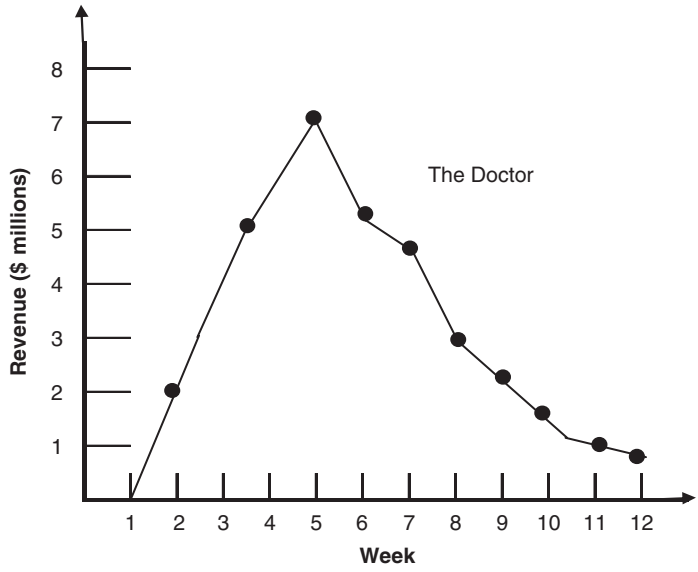

Fig. 3. Weekly Box Office Revenues for a Movie.

it is evident that the revenues for the movie grew till it reached its peak in week 4 and then declined gradually. Obviously, much of the revenue was generated through word-of-mouth influence, indicating that imitation factor dominates the innovation factor (i.e., $q > p$). We evaluate the three methods discussed in the previous section to determine the Bass model parameters and its efficacy in forecasting the movie revenues (Anderson et al., 2005).

RESULTS AND CONCLUSIONS

We used Lingo10 and MS Excel to calculate the Bass model parameters. Our results shown in Tables 1 and 2 indicate that all of the aforementioned techniques of the Bass model parameter estimation are comparable; however, the technique, OLS regression, has a slight edge over the other two techniques. The graph shown in Fig. 4 shows the actual and forecast trends for all the three methods.

The NLP method overestimates the forecast and hence higher error rates, whereas the LAD method underestimates the actual revenues and has error rates larger than the NLP. The OLS regression forecast closely follows the actual revenues and gives the best estimates of the revenues as evident from the low forecast errors. We can conclude that the NLP acts as an upper bound, whereas the LAD acts as a lower bound for the forecast estimates.

Table 1. Comparison of the Three Methods: Nonlinear Programming, Least Absolute Deviation, and OLS Regression.

Week	Weekly Sales ($ millions)	Forecast ($ millions)		
		NLP	LAD	LR
1	0.1	4.1	0.1	2.6
2	3.0	4.1	0.1	2.6
3	5.2	5.4	0.9	3.7
4	7.0	7.0	1.9	5.1
5	5.3	7.6	2.6	5.7
6	4.9	7.0	2.7	5.2
7	3.0	5.6	2.3	4.1
8	2.4	4.3	1.9	3.0
9	1.9	3.1	1.5	2.0
10	1.3	2.0	1.1	1.1
11	0.8	1.1	0.8	0.4
12	0.6	0.6	0.6	0.0

Table 2. Comparison of the Forecast Errors for the Three Methods: Nonlinear Programming, Least Absolute Deviation, and OLS Regression.

Week	Weekly Sales ($ millions)	Forecast Errors		
		NLP	LAD	LR
1	0.1	4.0	0.0	2.5
2	3.0	1.1	−2.9	−0.4
3	5.2	0.2	−4.3	−1.5
4	7.0	0.0	−5.1	−1.9
5	5.3	2.4	−2.6	0.4
6	4.9	2.1	−2.2	0.3
7	3.0	2.6	−0.7	1.1
8	2.4	1.9	−0.5	0.6
9	1.9	1.2	−0.4	0.1
10	1.3	0.7	−0.2	−0.2
11	0.8	0.3	0.0	−0.4
12	0.6	0.0	0.0	−0.6

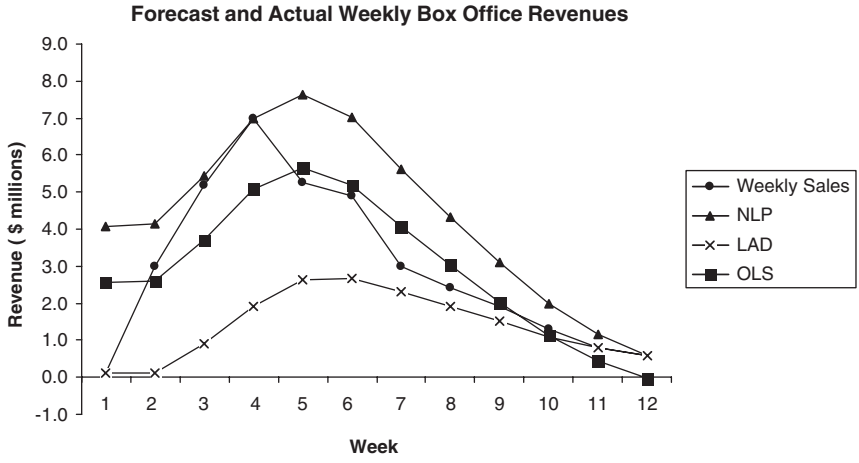

Fig. 4. Graph Showing Actual Revenues and the Forecast Revenues for the Three Methods.

The Bass model gives an appealing method to explain the diffusion of a new product in the absence of historical data. With rising product development, planning and product launch costs, the Bass model gives a good initial estimate of the forecasts. The model can be used for long-term

forecasting of the adoption of an innovation. In this chapter, we provided a comparative evaluation of the three methods of calculating the Bass model parameters and its estimation and use in forecasting applications. We found that the OLS regression method performs better in estimating the model parameters and hence in forecasting the revenues.

REFERENCES

Akinola, A. (1986). An application of the Bass model in the analysis of diffusion of cocospraying chemicals among Nigerian cocoa farmers. *Journal of Agricultural Economics*, *37*(3), 395–404.

Anderson, D. R., Sweeney, D. J., Williams, T. A., & Martin, K. (2005). *An introduction to management science* (12th ed.). Mason, OH: Thomson South-Western.

Bass, F. (1969). A new product growth model for consumer durables. *Management Science*, *15*, 215–227.

Bass, F. (1993). The future of research in marketing: Marketing science. *Journal of Marketing Research*, *30*, 1–6.

Dodds, W. (1973). An application of the Bass model in long term new product forecasting. *Journal of Marketing Research*, *10*, 308–311.

Fourt, L. A., & Woodlock, J. W. (1960). Early prediction of market success for grocery products. *Journal of Marketing*, *25*, 31–38.

Kalish, S., & Lilien, G. L. (1986). A market entry timing model for new technologies. *Management Science*, *32*, 194–205.

Lancaster, G. A., & Wright, G. (1983). Forecasting the future of video using a diffusion model. *European Journal of Marketing*, *17*(2), 70–79.

Lawrence, K. D., & Geurts, M. (1984). Converging conflicting forecasting parameters in forecasting durable new product sales. *European Journal of Operational Research*, *16*(1), 42–47.

Lawrence, K. D., & Lawton, W. H. (1981). Applications of diffusion models: Some empirical results. In: Y. Wind, V. Mahajan & R. Cardozo (Eds), *New product forecasting* (pp. 525–541). Lexington, MA: Lexington Books.

Lawton, S. B., & Lawton, W. H. (1979). An autocatalytic model for the diffusion of educational innovations. *Educational Administration Quarterly*, *15*, 19–53.

Lilien, G. L., & Rangaswamy, A. (2006). *Marketing engineering* (2nd ed.). New Bern, NC: Trafford Publishing.

Mahajan, V., & Muller, E. (1979). Innovation diffusion and new product growth models in marketing. *Journal of Marketing*, *43*, 55–68.

Mansfield, E. (1961). Technical change and the rate of imitation. *Econometrica*, *29*, 741–766.

Nevers, J. V. (1972). Extensions of a new product model. *Sloan Management Review*, *13*, 78–79.

Rogers, E. M. (1962). *Diffusion of innovations* (1st ed.). London: The Free Press.

Tigert, D., & Farivar, B. (1981). The Bass new product growth model: A sensitivity analysis for a high technology product. *Journal of Marketing*, *45*, 81–90.

Vanderbei, R. J. (2007). *Linear programming: Foundations and extensions* (3rd ed.). New York, NY: Springer.

This page intentionally left blank

THE USE OF A FLEXIBLE DIFFUSION MODEL FOR FORECASTING NATIONAL-LEVEL MOBILE TELEPHONE AND INTERNET DIFFUSION

Kallol Bagchi, Peeter Kirs and Zaiyong Tang

INTRODUCTION

Much attention has been given to adoption and diffusion, defined as the degree of market penetration, of Information and Communications Technologies (ICT) in recent years (Carter, Jambulingam, Gupta, & Melone, 2001; Kiiski & Pohjola, 2002; Milner, 2003; Benhabib & Spiegel, 2005). The theory of diffusion of innovations considers how a new idea spreads throughout the market over time. The ability to accurately predict new product diffusion is of concern to designers, marketers, managers, and researchers alike. However, although the diffusion process of new products is generally accepted as following an s-curve pattern, where diffusion starts slowly, grows exponentially, peaks, and then declines (as shown in Fig. 1), there is considerable disagreement about what factors affect diffusion and how to measure diffusion rates (Bagchi, Kirs, & Lopez, 2008).

Advances in Business and Management Forecasting, Volume 6, 93–107

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006007

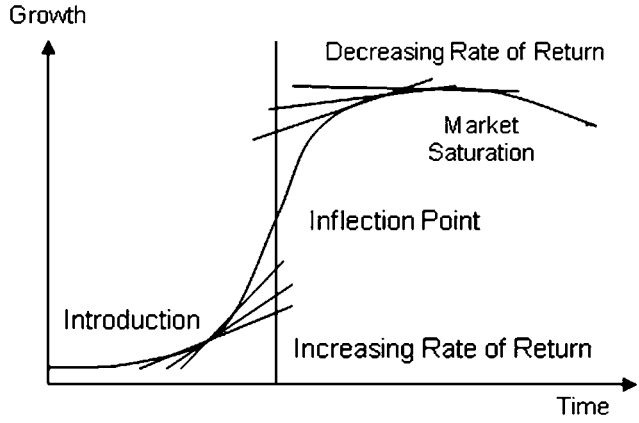

Fig. 1. New Product Diffusion.

Diffusion of technology products varies considerably across nations. For example, mobile phone growth has been spectacular in nations in general but particularly in nations such as India and China where usage increased from 3.53 and 65.82 per 1,000 in 2000 to 214.77 and 414.62 per 1,000 in 2007, respectively. Worldwide, the pattern of diffusion varies greatly. Mobile phone penetration for selected nations, or groups of nations, is given in Fig. 2; Internet penetration is given in Fig. 3. As the figures show, the level of diffusion, as well as the rate of diffusion (the speed at which penetration occurs), can greatly vary between regions.

Forecasting models are aimed at predicting future observations with a high degree of accuracy (Shmueli & Koppius, 2008). Diffusion models are used to forecast new product adoptions (Mahajan, Muller, & Bass, 1995) and the emphasis is on predicting the ultimate level of penetration (saturation) and the rate of approach to saturation. Diffusion models can also be considered a type of explanatory model, and have been used to clarify differences in national technology adoptions, based on estimated parameter values. Diffusion models thus not only yield good forecasting estimates; they generate parameter values, which can be used to explain difference in diffusion patterns.

Diffusion modeling studies try to explain and analyze patterns of innovation diffusion, usually over time and across a population of potential adopters, and forecast diffusion levels and rates. The emphasis is on predicting the ultimate level of penetration and the rate of approach to saturation. Observations of diffusion or percentages of diffusion are

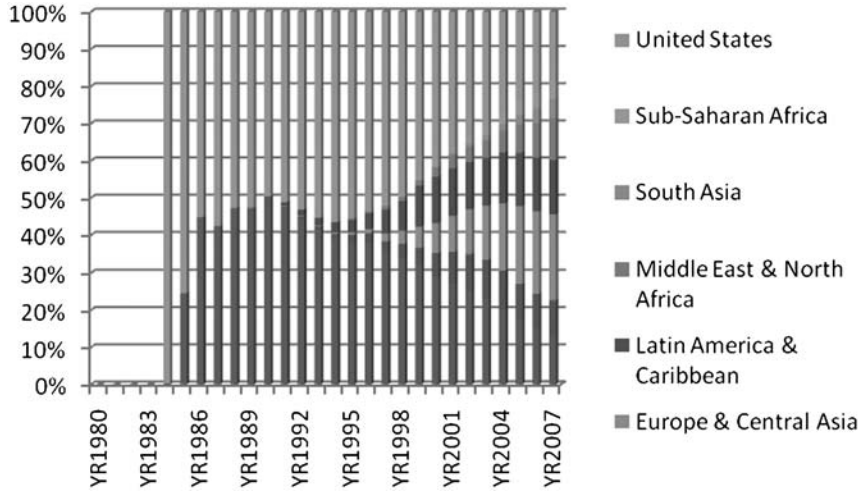

Fig. 2. Mobile Diffusion per 100 for a Few Regions.

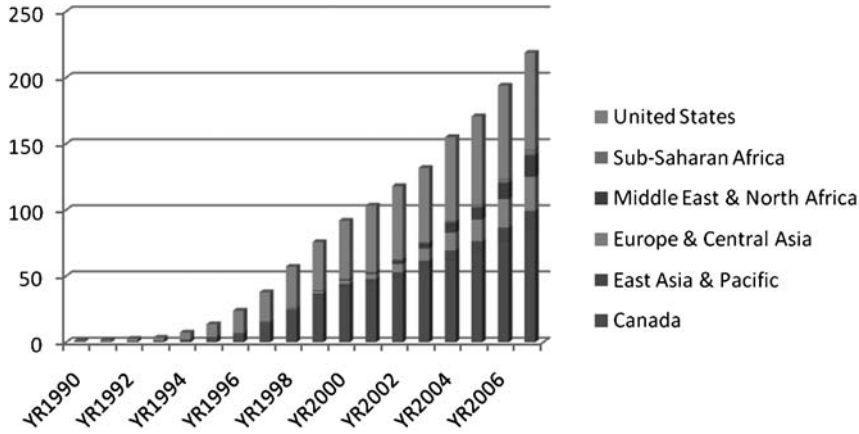

Fig. 3. Internet Diffusion per 100 for a Few Regions.

typically put in the form of a time series and fitted to some functional form, for example, the external influence model, internal influence models such as the logistic, Gompertz, or mixed influence models (Mahajan & Peterson, 1978).

Diffusion models with time varying parameters, such as the Von Bertalanffy (BERT) model (Bertalanffy, 1957) are called flexible models, and have previously been used to examine the diffusion process of an innovation (Hu, Saunders, & Gebelt, 1997). Incorporating time-variation in the parameters (b and θ, to be discussed later) of a diffusion model usually yields better estimates than time invariant models (Easingwood, 1989; Radas, 2005). By incorporating prior estimates of unknown parameters and updating initial estimates as new data become available, time-varying estimation procedures often can provide better early forecasts (Xie, Song, Sirbu, & Qiong Wang, 1997).

The role of ICT in economic development of a nation has been studied previously, in both developed and developing nations (Wong, 2002; Sridhar Kala, & Sridhar, 2003). It is generally agreed that investment in ICT positively influences the economic growth of a nation by reducing transaction costs and increasing output for firms in various sectors of the economy, which in turn promotes more investments in ICT (Roller & Waverman, 2001). In the present study, the diffusion of two technologies, mobile and the Internet, are investigated across different nations. These two technologies could be crucial for a country's economic growth and are at different stages in the product development life cycle. The Internet has become a new standard outlet for doing business in developed nations. In many developed nations, mobile phone growth has already exceeded 1 per user (The World Bank database, 2008). Expanding mobile networks brings in more customers, better foreign investments, and increase revenues for the government. According to a McKinsey & Company study, raising wireless penetration by 10% points can lead to an increase in gross domestic product (GDP) of about 0.5%, or around $12 billion for an economy the size of China (Dalka, 2007). Accurate forecasts of these technologies are, therefore, important to managers and government policy makers. Inaccurate forecasts in mobile handset as well as network equipment market can cost millions of dollars in losses (Wenrong, Xie, & Tsui, 2006).

It is the aim of this study to attempt to answer the following questions:

- How well does a flexible diffusion model such as BERT forecast the diffusion of two ICTs, the Internet and mobile, in different nations? In particular, how do these forecasts compare with traditional and nontraditional forecasting schemes?
- What parameter values for b and θ of the BERT model can be obtained for mobile phone and Internet diffusion in various nations? Are these values different for developing and developed nations for both of the products? For the two ICT products?

- What kind of analysis/interpretation we can make from these forecasts, which otherwise would be difficult to do with traditional forecasting models? For example, can the results be used to see whether decisions can be made to introduce a new IT product in different nations?

NATIONAL DIFFUSION OF ICT

It is generally accepted that national ICT diffusions can, in large part, be explained by country-specific factors such as demographic, economic, regulatory, infrastructural, educational, and affordability factors (Bagchi et al., 2008). In the same study, the authors also found that the impact of price decreases was positive and significant in all regions and for both telephone and mobile phone diffusion. Dedrick, Goodman, and Kraemer (1995) also used nation-specific analysis to investigate the factors responsible for the differences in ICT diffusion among nine culturally and geographically diverse small, developed nations. They concluded that the level of economic development, basic education system, infrastructure, investment in ICT, and government policies were reasons for different ICT diffusions among the countries. Watson and Myers (2001) used the Ein-Dor, Myers, and Raman (1997) approach to investigate the ICT industry and ICT diffusion in Finland and New Zealand using 1998 data. They concluded that although the two countries were similar in many ways, Finland's ICT diffusion outpaced New Zealand's due to government promotion of ICT, research and development in private sector, and an ICT-based education system.

In a previous paper (Bagchi, Kirs, & Udo, 2006), the authors contrasted mobile phone, PC, and Internet diffusion levels and rates between developed and developing nations and found a number of differences in factors responsible for such diffusions between the two groups. For developed nations the most significant factors influencing ICT diffusion rates were human development, urbanization, and institutional factors. Human development, as conceived by the United Nations Development Programme (UNDP, 1990), is a composite index of normalized measures of life expectancy, literacy, educational attainment, and GDP per capita for countries worldwide. Institutional factors, or the Economic Freedom of the World (EFW) index, is a composite index containing 38 components designed to measure the degree to which a nation's institutions and policies are consistent with voluntary exchange, protection of property rights, open markets, and minimal regulation of economic activity (Gwartney & Lawson, 2003;

Walker, 1988). In contrast, for developing countries, the significant factors influencing ICT diffusion rates included ICT infrastructure, human development, and income disparity.

Although it can be anticipated that developed countries can initially best take advantage of innovative technologies due to purchasing power, the assumption that established technology diffusion rate in developed countries is always faster than in lesser developed countries may be misplaced, at least in a later stage of diffusion. Gruber and Verboven (1998) showed that later adopting countries have faster diffusion rates in some stages, in accordance with diffusion theory, implying that developing countries can take advantage of a "leap-frog" affect (Davison, Vogel, Harris, & Jones, 2000) by applying an "investment-based growth strategy" (Acemoglu, Aghion, & Zilibotti, 2002). As national scenarios for technology diffusion widely vary, diffusion models such as flexible ones that assume asymmetric diffusion patterns could be more appropriate for modeling technological diffusions in different nations. The parameters of the diffusion model can be expected to have significantly different values for developing and developed nation diffusion scenarios.

DIFFUSION MEASUREMENT

Although there are a number of diffusion model estimation procedures, Mahajan, Muller, and Bass (1990) offer two basic model classes: *Time-Invariant Estimation Procedures* (TIEP) and *Time-Varying Estimation Procedures* (TVEP). In TIEP, the output does not depend explicitly on time and includes conventional estimation methods such as ordinary least square (OLS) (Bass, 1969), maximum likelihood estimation (MLE) (Schmittlein & Mahajan, 1982), and nonlinear least squares (NLS) (Srinivasan & Mason, 1986). However, according to Xie et al. (1997), there are two common limitations inherent in TIEP. First, to obtain stable and robust parameter estimates, TIEP often require data to include the peak sales (Mahajan et al., 1990). TIEP are also not helpful in forecasting a new product diffusion since they require observable data to be collected over time; by the time sufficient data have been collected, it may be too late for adequate forecasting or planning. Second, TIEP can be applied only to a discrete form of a diffusion model or to a solution to a diffusion model. These discrete forms can often result in biased and high variance estimates.

TVEP have been introduced to overcome some of these limitations of TIEP (Mahajan et al., 1990). TVEPs start with a prior estimate of unknown

parameters in a diffusion model and update the estimates as additional data become available. Time-varying estimation procedures in the marketing science literature include the BERT model (Bertalanffy, 1957), the adaptive filter procedure (Bretschneider & Mahajan, 1980), the Hierarchical Bayesian (Lenk & Rao, 1990), and nonlinear learning rate adaptation (Bousson, 2007). Flexible models fall within this class. As Xie et al. (1997) note, diffusion models should have at least two desirable properties. They should facilitate forecasts early in the product cycle, when only a few observations are available, and should provide a systematic way of incorporating prior information about the likely values of model parameters and an updating formula to upgrade the initial estimates as additional data become available. Second, they should be expressed as a differential equation and should require neither a discrete analog (i.e., it is not required that a continuous differential equation be rewritten as a discrete time equation in a way that introduces a time interval bias) nor an analytic solution to the equation.

METHODOLOGY

Flexible diffusion models have the added advantage over other traditional diffusion models (internal, external, or mixed) in that they do not assume that the point of inflection at which the maximum diffusion is reached has to be at a point when 50% or less adoption has taken place. The BERT diffusion model is an example of a flexible diffusion model.

The functional form of the BERT model is given by

$$\frac{dF(t)}{dt} = \frac{b}{(1-\theta)}[F^{\theta}(t)][m^{1-\theta}(t) - F^{1-\theta}(t)]$$

where $F(t)$ is the number of adopters at time $T = t$; m the potential number of adopters at time $T = t$; and $b > 0$ and θ (th) > 0 are model parameters that determine the nature of the model.

It can be observed that when

- $\theta = 0$, the model reduces to external influence model,
- $\theta = 1$, the model reduces to Gompertz internal influence model,
- $\theta = 2$, the model reduces to internal influence model with $q = b/m$,
- θ has other values, the model reduces to mixed influence model.

The parameters to be evaluated for the model are m, θ, and b.

TEST SCHEMES

The methodology consisted of running nonlinearly the flexible model in Excel Solver and running standard forecasting techniques such as exponential smoothing (ES) and moving averages (MA) using SPSS (Kendrick, Mercado, & Amman, 2006). The parameter values of the BERT diffusion model were obtained for developing and developed nations after running the Excel Solver, and suitable statistical tests were conducted for significant differences in values across the set of nations.

For forecasting purposes, separate year data (2007) were used. For in-sample and out-of-sample forecast comparisons, six regions and two technologies were selected. The selection was based on sum of squares of errors (SSE) values of the BERT model fits and two regions each were selected for best, worst, and middle-level values of SSEs for each technology. Two standard forecasting procedures, ES and MA were selected. The ES procedure computes forecasts for time series data using exponentially weighted averages. The number of periods in the MA procedure was selected to remove seasonal effect. To keep calculations manageable, for ES, a middle value of smoothing constant/damping factor (0.30) was selected; in practice, the values of a smoothing constant typically range from 0.05 to 0.60 (Statistix for Windows, 1996). For MA, the number of periods selected was 3. The data from 2007 was used for out-of-sample forecasting and squared difference was computed (DiffSE07). For in-sample forecasting, the sum of squares of all forecast differences from different years was computed (DiffSE90-06) for comparison among various forecasting methods.

DATA

All data was obtained from The World Bank database (2008). The data consists of time series data of yearly Internet and the mobile diffusions per 100 residents from 24 groups of nations/regions/nations. All the definitions of groups can be found in The World Bank database (2008). The groups are of two types – region- and economy-based. Although, there is some overlap of nations in various groups, the groups are by themselves representative of developing and developed sets of nations. Large nations such as India and China, as well as developed nations such as the U.S, the U.K., and Australia, were also included to maintain representation of the entire world, and because these two technologies have been extensively adopted in these nations. Internet data were from 1990 to 2006 and the mobile data were from 1987 to 2006.

RESULTS

The results of the model are shown in Table 1. The model fits in general are good with a few exceptions. For two nations, high income non-Organisation for Economic Co-operation and Development (OECD) and Australia, the BERT model for Internet diffusion underestimated the actual diffusion scenario. Except for a few nations, model fits for mobile (high SSEs, mean = 50.88) are poorer than the Internet (SSE mean = 26.77). The pairwise *t*-test rejects equality of mean values (with $t = -1.345$, $p = 0.193 > 0.05$). Pairwise *t*-tests of θ ($t = 1.08$, $p > 0.05$) and b ($t = 0.929$, $p > 0.05$) of mobile and the Internet showed that these values are different for the Internet and mobile diffusions, suggesting that these technologies have diffused differently for the same nation/nation groups.

Table 1. Results of Bertalanffy Model Estimates.

Nations	Internet				Mobile			
	m	θ	b	SSE	m	θ	b	SSE
Developed regions/groups/nations								
Upper middle income	104.713	1.122	0.136	1.667	295.106	1.382	0.213	20.870
The US	120.524	0.685	0.109	100.890	394.088	0.735	0.050	15.186
The UK	129. 399	0.875	0.122	81.783	115.436	1.621	0.468	353.545
OECD	235.181	1.003	0.082	3.480	102.716	1.329	0.303	60.917
High income non-OECD	45.083	1.349	0.342	7.157	106.220	1.553	0.396	106.829
Europe	241.135	1.196	0.138	2.695	364.735	1.757	0.362	14.720
Euro region	196.472	0.772	0.072	107.768	102.639	1.802	0.560	177.162
Canada	196.647	0.717	0.082	202.369	137.704	0.862	0.093	12.796
Australia	48.817	1.701	0.671	52.593	168.903	0.996	0.142	76.172
Developing regions/groups/nations								
Middle East	45.538	3.794	0.979	0.623	154.376	5.243	1.636	16.610
Middle income	43.436	1.452	0.238	0.869	122.913	1.605	0.283	1.968
Sub Saharan Africa	30.000	7.228	2.182	0.623	223.586	3.701	1.005	1.631
South East Asia	116.280	4.080	1.210	2.420	223.586	4.079	1.100	22.170
The World	44.043	1.012	0.135	0.927	125.4 34	1.186	0.151	8.046
Low and middle	44.166	1.595	0.274	0.829	122.862	1.655	0.292	1.454
Least developed	30.000	9.522	2.388	6.732	223.586	4.258	1.076	7.423
Low income	30.000	7.295	2.133	0.149	223.586	4.227	1.128	17.500
Latin	235.181	1.003	0.082	3.480	250.650	1.270	0.166	22.303
India	70.081	6.828	2.350	5.405	76.461	5.658	1.666	17.080
Heavy indebted	18.303	5.489	1.390	0.178	223.553	3.479	0.846	4.415
East Asia	61.685	1.142	0.126	2.540	42.778	2.195	0.556	3.884
China	104.755	1.090	0.103	3.854	264.750	1.920	0.490	156.767

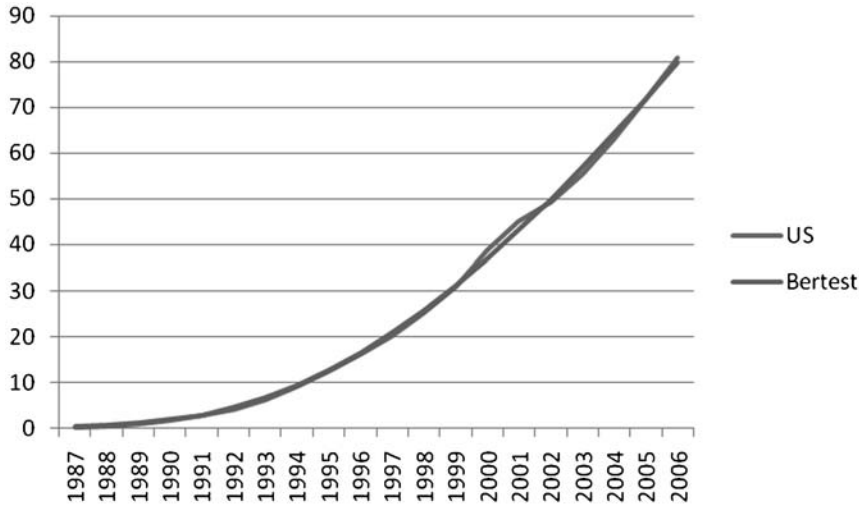

Fig. 4. Estimation of Von Bertalanffy Model for the US Mobile Diffusion.

Figs. 4 and 5 show the actual time series for the United States and the fit obtained from the BERT model. It is obvious that the model fit is close, though the SSE values obtained for the United States were not among the lowest.

The results for the Internet and mobile forecasts are shown in Tables 2 and 3. For the Internet, the six groups selected were low income, heavily indebted (low SSEs), Australia and middle income (medium SSEs), and Canada and Euro region (high SSEs). For four regions, BERT predictions proved to be superior to those of MA and ES, for both DiffSE07 and DiffSE90-06. For two other regions, Australia and Euro region, BERT model predictions were not superior for DiffSE07, although it performed better in DiffSE90-06.

For mobile telephone, the six groups selected were low and middle income, middle income (low SSEs), OECD and Middle East (medium SSEs), and the United Kingdom and Euro region (high SSEs). Again, for four regions, BERT predictions proved to be superior to those of MA and ES, for both DiffSE07 and DiffSE90-06 (see Table 3). For two other regions, the United Kingdom and Euro region, BERT predictions were not superior for DiffSE07, although it performed better in DiffSE90-06. Thus, as far as the Internet and mobile diffusion forecasting were concerned, BERT predictions were not inferior with respect to DiffSE07 or DiffSE90-06.

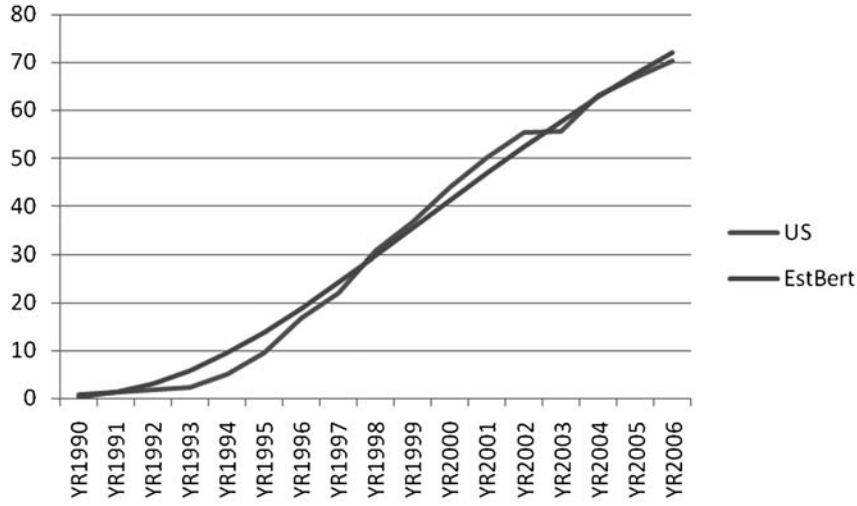

Fig. 5. Estimation of Von Bertalanffy Model for the US Internet Diffusion.

Table 2. Internet Forecasting of Six Regions (In Sample and Out-Of-Sample Forecasts).

		Internet Forecasting					
	Nation	Low income	Heavily indebted	Australia	Middle income	Canada	Euro region
Year 2007	Actual data	5.214	2.752	53.280	17.750	84.910	51.467
	MAPrediction	3.993	1.470	50.350	13.285	76.523	48.025
	ESPrediction	3.478	1.033	49.770	11.494	73.269	46.700
	BertPrediction	4.461	2.633	48.670	15.460	83.720	57.582
DiffSE07	MAPrediction	1.491	1.644	8.617	19.936	70.342	11.846
	ESPrediction	3.014	2.955	12.366	39.138	135.513	22.722
	BertPrediction	0.567	0.014	21.390	5.244	1.416	37.396
DiffSSE90-06	MAPrediction	2.531	2.856	337.309	19.973	515.269	226.580
	ESPrediction	4.982	0.485	656.758	40.350	1,022.140	453.030
	BertPrediction	0.149	0.178	73.982	0.869	202.370	107.768

Notes: DiffSE07, squared difference between actual data and prediction model for 2007; DiffSSE90-06, sum of squared difference between actual data and prediction model for years 1990–2006.

Table 3. Mobile Forecasting of six Regions (In Sample and Out-Of-Sample Forecasts).

		Mobile Forecasting					
	Nation	Middle East and North Africa	OECD	Low and mid income	Middle income	UK	Euro region
Year 2007	Actual data	50.7	94.995	41.503	46.594	117.950	107.685
	MAPrediction	35.940	90.702	32.877	34.858	113.932	103.994
	ESPrediction	29.950	89.018	29.424	38.228	112.072	102.609
	BertPrediction	48.333	93.254	41.398	47.769	112.179	100.999
DiffSE07	MAPrediction	217.869	18.430	74.408	137.734	16.144	13.623
	ESPrediction	430.500	35.725	145.902	69.990	34.551	25.766
	BertPrediction	5.603	3.031	0.011	1.381	33.304	44.703
DiffSSE90-06	MAPrediction	213.632	651.179	139.834	187.477	1,451.874	1,189.536
	ESPrediction	414.820	1,296.889	275.922	369.475	2,818.429	2,330.951
	BertPrediction	16.614	61.674	1.454	1.968	353.545	177.162

Notes: DiffSE07, squared difference between actual data and prediction model for 2007; DiffSSE90-06, sum of squared difference between actual data and prediction model for years 1990–2006.

DISCUSSIONS AND CONCLUSION

This research provides some answers to the research questions posed in the ''Introduction'' section. Traditional diffusion models can be inadequate to model the technology product diffusion in various nations. The traditional diffusion models (internal, external, and mixed) cater to standard distribution pattern of product diffusion over time. Diffusion models for technology diffusion in various nations must be more flexible to capture the variety of different patterns in different nations that may exist, depending on the various national conditions prevalent during the period of diffusion.

The results showed that a flexible diffusion model such as BERT can indeed model the diffusion of the ICT products more realistically in both developed and developing nations. The model predictions, in most cases, were better than traditional forecasting schemes such as ES or MA. For the two technologies, 4 out of 6 out-of-sample BERT forecasts emerged as better. Although not shown in this research, the BERT model predictions were, in most cases, also better than those from the Bass model (Bass, 1969).

The parameter values were also different for these two sets of nation groups, developing and developed. Table 1 shows that developing nations in

general, had greater values in b and θ. One interpretation of this result could be that the diffusion pattern of these technologies in developing nations requires more word-of-mouth and promotional effort as not many people are ready to adopt these products due to differences in developing nation's social, institutional, and economic conditions. At the same time, in many developing nations the potential for high diffusion of these products exists (as the values of m show).

The parameter values also showed that for each nation or nation-group, mobile diffusion has progressed differently than the Internet diffusion. This is because, in general, Internet subscriptions per 100 individuals have been slower than mobile diffusion per 100. Price could be one of the major factors for such differences of diffusions in a same nation/set of nations (Bagchi et al., 2008). To give some concrete examples of how the diffusions have differed, consider the diffusion of two technologies of two nation sets at extreme economic ends: heavily indebted poor nations and high income nations. From the period 1990 to 2007, mobile diffusion has progressed in the heavily indebted poor nations and high-income nations from base values of 0 and 1.16, respectively, to 15.77 and 96.99, respectively. For the Internet, during the same period, high income nations registered a much slower increase from 0.28 to 63.53 and heavily indebted poor nations registered a similar slower increase from 0 to 2.75 (The World Bank database, 2008). Although a few situations can be found where the Internet diffused faster than mobile phones during the same time period (Canada is an example), for most nations the reverse has been the case.

The results from the BERT model indicate that it can be used effectively in predicting the diffusion of these technologies in different nations and the model may yield better results in many cases than the traditional ES or MA schemes or the standard Bass model of diffusion. For a successful diffusion of the product in any nation, a manager needs to additionally investigate the relationship of model parameter values to various types of existing social, institutional, and economic factors in a nation. Future research can also relate the model parameters b and θ more specifically to business situations (Radas, 2005).

REFERENCES

Acemoglu, D., Aghion, P., & Zilibotti, F. (2002). *Distance to frontier, selection, and economic growth*. NBER Working Paper 9066, National Bureau of Economic Research, Inc.

Bagchi, K., Kirs, P. J., & Lopez, F. (2008). The impact of price on cell phone and telephone diffusion in four groups of nations. *Information & Management*, *45*, 183–193.

Bagchi, K., Kirs, P. J., & Udo, G. (2006). A comparison of factors impacting ICT growth rates in developing and industrialized countries. In: E. Trauth, D. Howcroft, T. Butler, B. Fitzgerald, & J. DeGross (Eds), *IFIP International Federation for Information Processing, Volume 208*. Social inclusion: Societal and organizational implications for information systems (pp. 37–50). Boston: Springer.

Bass, F. M. (1969). A new product growth model for consumer durable. *Management Science*, *15*, 215–227.

Benhabib, J., & Spiegel, M. M. (2005). Human capital and technology diffusion. *Handbook of Economic Growth*, *1*(Part A), 935–966.

Bertalanffy, L. Von. (1957). Quantitative laws in metabolism and growth. *Quarterly Review of Biology*, *32*, 217–231.

Bousson, K. (2007). Time-varying parameter estimation with application to trajectory tracking. *Aircraft Engineering and Aerospace Technology*, *79*(4), 406–413.

Bretschneider, S. I., & Mahajan, V. (1980). Adaptive technological substitution models. *Technological Forecasting & Social Change*, *18*(2), 129–139.

Carter, F. J., Jambulingam, T., Gupta, V. K., & Melone, N. (2001). Technological innovations: A framework for communicating diffusion effects. *Information & Management*, *5*, 277–287.

Dalka, D. (2007). *Mobile phone usage creates economic growth in developing nations* (Available at http://www.daviddalka.com/createvalue/2007/01/28/mobile-phone-usage-creates-economic-growth-in-developing-nations/).

Davison, R., Vogel, D., Harris, R., & Jones, N. (2000). Technology leapfrogging in developing countries: An inevitable luxury? *The Electronic Journal on Information Systems in Developing Countries*, *5*, 1–10.

Dedrick, J. L., Goodman, S. E., & Kraemer, K. L. (1995). Little engines that could: Computing in small energetic countries. *Communications of ACM*, *38*(5), 21–26.

Easingwood, C. (1989). An analogical approach to the long-term forecasting of major new product sales. *International Journal of Forecasting*, *5*, 69–82.

Ein-Dor, P., Myers, M. D., & Raman, K. S. (1997). Information technology in three small developed countries. *Journal of Management Information Systems*, *13*(4), 61–89.

Gruber, H., & Verboven, F. (1998). *The diffusion of mobile telecommunications services in the European union*. Discussion Paper 138. Center for Economic Research, Tilburg University.

Gwartney, J., & Lawson, R. (2003). *Economic freedom of the world: 2003 annual report*. Vancouver: The Fraser Institute.

Hu, Q., Saunders, C., & Gebelt, M. (1997). Research report. Diffusion of information systems outsourcing: A reevaluation of influence sources. *Information Systems Research*, *8*, 288–301.

Kendrick, D., Mercado, P., & Amman, H. (2006). *Computational economics*. Princeton, NJ: Princeton University Press.

Kiiski, S., & Pohjola, M. (2002). Cross country diffusion of the internet. *Information Economics and Policy*, *14*(2), 297–310.

Lenk, P. J., & Rao, A. G. (1990). New models from old: Forecasting product adoption by hierarchical Bayes procedures. *Marketing Science*, *9*(1), 42–53.

Mahajan, V., Muller, E., & Bass, F. M. (1990). New product diffusion models in marketing: A review and directions for re-search. *Journal of Marketing*, *54*, 1–26.

Mahajan, V., Muller, E., & Bass, F. M. (1995). Diffusion of new products: Empirical generalizations and managerial uses. *Marketing Science*, *14*(3), G79–G88.

Mahajan, V., & Peterson, R. (1978) Models for innovation diffusion. *Sage University Paper series on Quantitative Applications in the Social Sciences* (2nd ed.). Beverly Hills and London: Sage.

Milner, H. V. (2003). *The global spread of the internet: The role of international diffusion pressures in technology adoption*. New York: Columbia University.

Radas, S. (2005). Diffusion models in marketing: How to incorporate the effect of external influence? *Economic Trends and Economic Policy*, *15*(105), 31–51.

Roller, L. H., & Waverman, L. (2001). Telecommunications infrastructure and economic development: A simultaneous approach. *American Economic Review*, *91*(4), 909–923.

Schmittlein, D. C., & Mahajan, V. (1982). Maximum likelihood estimation for an innovation diffusion model of new product acceptance. *Marketing Science*, *I*, 57–78.

Shmueli, G., & Koppius, O. (2008). *Contrasting predictive and explanatory modeling in IS research*. Robert H. Smith School Research Paper no. RHS 06-058. Available at http://ssrn.com/abstract = 1112893

Sridhar Kala, S., & Sridhar, V. (2003). The effect of telecommuting on suburbanisation: Empirical evidence. *Journal of Regional Analysis and Policy*, *33*(1), 1–25.

Srinivasan, V., & Mason, C. (1986). Nonlinear least squares estimation of new product diffusion models. *Marketing Science*, *5*, 169–178.

Statistix for Windows. (1996). Analytical Software, ISBN 1-881789-04-7.

The World Bank database. (2008). Available at http://www.devdata.worldbank.org

UNDP. (1990). World Development Report 1990. Oxford: UNDP.

Walker, M. A. (Ed.) (1988). *Freedom, democracy, and economic welfare*. Vancouver: The Fraser Institute.

Watson, R., & Myers, M. D. (2001). IT industry success in small countries: The case of Finland and New Zealand. *Journal of Global Information Management*, *9*(2), 4–14.

Wenrong, W., Xie, M., Tsui, K. (2006). Forecasting of mobile subscriptions in Asia pacific using bass diffusion model. *Proceedings 2006 IEEE International Conference on Management of Innovation and Technology*, Singapore (pp. 300–303).

Wong, P.-K. (2002). ICT production and diffusion in Asia: Digital dividends or digital divide? *Information Economics and Policy*, *14*, 167–187.

Xie, J. X., Song, M., Sirbu, M., & Qiong Wang, Q. (1997). Kalman filter estimation of new product diffusion models. *Journal of Marketing Research*, *34*(3), 378–393.

This page intentionally left blank

FORECASTING HOUSEHOLD RESPONSE IN DATABASE MARKETING: A LATENT TRAIT APPROACH

Eddie Rhee and Gary J. Russell

ABSTRACT

Database marketers often select households for individual marketing contacts using information on past purchase behavior. One of the most common methods, known as RFM variables approach, ranks households according to three criteria: the recency of the latest purchase event, the long-run frequency of purchases, and the cumulative dollar expenditure. We argue that RFM variables approach is an indirect measure of the latent purchase propensity of the customer. In addition, the use of RFM information in targeting households creates major statistical problems (selection bias and RFM endogeneity) that complicate the calibration of forecasting models. Using a latent trait approach to capture a household's propensity to purchase a product, we construct a methodology that not only measures directly the latent propensity value of the customer, but also avoids the statistical limitations of the RFM variables approach. The result is a general household response forecasting and scoring approach that can be used on any database of customer transactions. We apply our

Advances in Business and Management Forecasting, Volume 6, 109–131

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006008

methodology to a database from a charitable organization and show that the forecasting accuracy of the new methodology improves upon the traditional RFM variables approach.

INTRODUCTION

Database marketing is an increasingly important aspect of the management of traditional catalog retailers (such as Lands' End) and e-commerce firms (such as Amazon.com). In database marketing, the manager has access to a household database detailing each interaction with the firm over a period of time. The task of the marketing manager is to use the database information to develop predictive models of household purchasing, and then to target segments of households for specific marketing programs (Winer, 2001; Berger & Nasr, 1998; Hughes, 1996).

Scoring Households Using RFM

Firms frequently use household purchase characteristics, collectively known as RFM variables, in selecting the best households for marketing solicitations (Hughes, 1996). Recency (R) is defined as the number of periods since the most recent purchase. Frequency (F) is defined as the total number of purchases. Monetary value (M) is defined as the dollar amount that the household has spent to date. Conceptually, RFM variables are used for forecasting because past purchase behavior is often a reliable guide to future purchase behavior (Schmid & Weber, 1997; Rossi, McCulloch, & Allenby, 1996). The predictive power of the three variables is traditionally known as having the rank order: recency is the best predictor, followed by frequency, and then monetary value (David Sheppard Associates, Inc., 1999). Forecasting the customer's response likelihood using RFM variables is widely accepted by database marketers as an easy and useful way of predicting behavior from a customer database.

RFM information is inserted into predictive models. For example, RFM values can be used as independent variables in a probit or logit response model. Additional procedures drawn from the data mining literature, such as decision trees and neural networks, can also be used to link RFM values to buying behavior (Berry & Linoff, 2000).

Statistical Problems Induced by RFM

The use of RFM in response modeling appears straightforward on the surface but important statistical problems arise.

The first problem is known as selection bias. Simply put, selection bias arises when the researcher uses a nonrandomly selected sample to estimate behavioral relationships (Heckman, 1979). If the firm selects households for mailings based on a nonrandom selection rule (such as the RFM variables), a study that only analyzes the selected households generates biased results. This bias arises from the fact that the researcher does not observe the behavior of nonselected households. Selection bias is a special type of missing data problem that can only be controlled by formally analyzing the way that the firm selects customers for marketing solicitations.

The second problem, known as RFM endogeneity, occurs when RFM values not only represent the past response behavior of the households, but also reflect the past selection decision of the firm. For instance, if a household is not selected to receive a marketing offer (and the household has no way to respond to the offer otherwise), the recency (the number of periods since the last purchase) will be larger and the frequency and the monetary value will be smaller, than the values of these same variables for a comparable household who received the solicitation. If the firm consistently ignores the household for any reason, the RFM values of this household will deteriorate regardless of the true propensity to respond. In formal statistical terms, it can be shown that RFM endogeneity yields incorrect parameter estimates in a predictive model due to unobserved correlations between the RFM variables and the error in the model (see, e.g., Davidson & MacKinnon, 1993).

The marketing science community has gradually begun to recognize these statistical problems. Industry standard procedures (such as RFM probit regression) and the early model of Bult and Wansbeek (1995) ignore these problems entirely. Jonker, Paap, and Frances (2000) addresses selection bias by relating RFM variables to both household selection and household response. However, because RFM values appear in the specification, issues of endogeneity are not addressed. Studies by Bitran and Mondschein (1996) and Gonul and Shi (1998) provide an approach to dealing with RFM endogeneity. By replacing observed RFM values with predicted values, these authors construct an instrumental variables methodology (see Davidson & MacKinnon, 1993) that corrects for potential parameter biases. In the applications discussed by these authors, households are able to buy products even if a marketing solicitation is not received. Accordingly, parameter biases due to selection bias are not relevant and are not addressed.

Latent Trait Scoring Model

This research builds on existing work by developing an approach to household scoring, which corrects for both selection bias and RFM endogeneity. In contrast to earlier studies, we assume that each household has a latent (unobserved) propensity to respond that cannot be adequately captured using RFM variables. Latent trait models have a long history in psychometric studies of psychological constructs such as verbal and quantitative ability (see, e.g., Lord & Novick, 1968; Fischer & Molenarr, 1995; Langeheine & Rost, 1988). These models have also found marketing science applications in survey research (Balasubramanian & Kamakura, 1989), coupon redemption (Bawa, Srinivasan, & Srivastava, 1997), and cross-selling of financial services (Kamakura, Ramaswami, & Srivastava, 1991). Although latent trait models can be regarded as a type of random coefficient heterogeneity model (Allenby & Rossi, 1999), they are best viewed as a method of measuring a psychological trait. In this research, we view household scoring as a research procedure designed to estimate a household's propensity to respond to marketing solicitations by the firm.

Our model is based on the assumption that both the firm's selection rule and the household's response behavior provide indirect indications of the household's latent propensity. The notion here is that the firm does *not* select households for mailings using either a census (all households) or a random process (probability sample of households). Instead, the firm selects households using some process that takes into account the likelihood that the household will respond favorably. We do not, however, assume that the selection process is necessarily optimal. The key advantage of this approach is generality: the researcher can estimate a household response model on existing databases in which the firm has attempted to optimize customer contact policy. As we show subsequently, we are able to measure each household's true propensity to respond and examine the effectiveness of the firm's current contact policy.

The remainder of the chapter is organized as follows. We first detail our new model, discussing the need to consider both household response and the firm's household selection rule simultaneously. We demonstrate that our latent trait specification can be formulated as a Hierarchical Bayes model and estimated using Monte Carlo simulation technologies. The new methodology is then applied in an analysis of the customer database of a nonprofit organization. We show that the model provides forecasts with a level of accuracy better than a benchmark RFM probit model. We conclude with a discussion of future research opportunities.

LATENT TRAIT MODEL OF RESPONSE PROPENSITY

The chapter begins by describing the structure of a general model of household choice behavior in a database marketing context. Instead of relying on RFM variables to measure the propensity of each household to respond to marketing solicitation, we assume the existence of a household-specific latent trait that impacts the household's probability of responding to a solicitation. This same latent variable is also assumed to impact the firm's likelihood of targeting the household. By developing the model in this manner, we correct for selection bias (if present) and avoid issues of RFM endogeneity. The result is a general model of household purchase behavior that takes into account potential limitations of the RFM variables approach.

Propensity to Respond

Propensity to respond is defined here as a household characteristic that reflects the household's inherent interest in the firm's product offering. We assume that this propensity has two components: a long-run component that varies only by household and a short-run component that varies over households and time. Implicitly, the long-run component accounts for heterogeneity in response across households, whereas the short-run component accounts for temporal variation in response propensity within households.

Let $\tau_{h,t}$ denote the propensity to respond of household h at time t. We define this construct according to the following equation:

$$\tau_{h,t} = \delta_1 \mu_h + \delta_2 \, Y_{R(h,t-1)} \tag{1}$$

where $Y_{R(h,t-1)} = 1$ if the household responded to the previous solicitation, and 0 otherwise. Furthermore, we assume that the long-run component is normally distributed across the household population as

$$\mu_h \sim N(X_h \gamma, 1.0) \tag{2}$$

where X_h denotes a set of demographic variables for household h.

This formulation has two key properties. First, $\tau_{h,t}$ changes over time depending on whether or not the household purchased at the previous time

point. Note that nonpurchase at time $t-1$ could be due to a rejection of the previous offer. Alternatively, the household may have not been given an opportunity to purchase the product. In our formulation, it is not necessary to distinguish between these two cases. Rather, similar to models of choice inertia found in grocery scanner data applications (Seetharaman & Chintagunta, 1998; Jeuland, 1979), we assume that the act of purchasing a product at one time point has an impact on future behavior, regardless of why the product was purchased. Note that the parameter on the lagged component (δ_2) can be either positive or negative. Thus, the impact of the short-term component can either enhance or diminish purchasing at the next period.

The normal distribution characterizing the long-run component is intended to allow for heterogeneity in response across households. Note that this distribution has a mean that depends on demographics and a variance set to one. Intuitively, this formulation states that the long-run component depends in part on demographics, and in part on other (unobserved) household characteristics. Setting the variance of this distribution to one can be accomplished without loss of generality. This restriction is necessary for model identification and does not impact the fit of the model.

Modeling the Firm's Targeting Decision

We assume that the firm is attempting to optimize its targeting policy using some information in the household database. This information may or may not be RFM variables. We assume that the firm's rule has some validity and is correlated to some extent with $\tau_{h,t}$ the household's propensity to respond. *We stress that this assumption does not imply that the firm actually observes* $\tau_{h,t}$. Rather, the expression displayed later for the firm's selection rule is simply a formal way of stating that households are not necessarily selected at random. During model estimation, the researcher learns the extent to which the current targeting policy is based on some knowledge of household's true propensity to respond. In the language of econometric theory, our model of the firm's targeting policy is a limited information specification – not a structural specification.

To model this process, we assume that the firm's decision to target a household depends on the attractiveness of the household to the firm. Define the attractiveness of a household h to the firm at time t as $U_{S(h,t)}$. We assume that the firm makes a product offer to this household ($Y_{S(h,t)} = 1$)

if $U_{S(h,t)}$ is greater than zero. Otherwise, $Y_{S(h,t)} = 0$. Hence, $U_{S(h,t)}$ is a latent variable that drives the observed targeting policy of the firm.

To complete the specification, we assume that the deterministic part of $U_{S(h,t)}$ is a linear function of the household's propensity to respond $\tau_{h,t}$ defined in Eq. (1). This leads to the model

$$U_{S(h,t)} = \alpha_0 + \alpha_1 \mu_h + \alpha_2 \, Y_{R(h,t-1)} + \varepsilon_{S\,h,\,t}, \quad \varepsilon_{S\,h,t} \sim N(0,1) \tag{3}$$

where the normally distributed error $\varepsilon_{S\,h,t} \sim N(0,1)$ has mean 0 and variance 1.

The assumption that the variance of the error is equal to 1 is necessary for model identification; it has no impact on model fit.

We again emphasize that this expression does not imply that the firm knows the household propensity to respond as measured by $\tau_{h,t}$. All that this expression states is that the attractiveness of a household to the firm is correlated to some extent with the household's propensity to respond. This specification of the household selection process is identical to that of a probit model for the binary variable $Y_{S(h,t)}$. Intuitively, this model allows for the possibility that households that are selected are likely to be better prospects for the firm.

Modeling Household Response

In an analogous fashion, we assume that the household's decision to respond to a product offer depends on the attractiveness (or utility) of the offering. Define the attractiveness of a marketing offering to household h at time t as $U_{R(h,t)}$. We assume that the household buys the product ($Y_{R(h,t)} = 1$) if $U_{R(h,t)}$ is greater than zero. Otherwise, $Y_{R(h,t)} = 0$. Hence, $U_{R(h,t)}$ is a latent variable that determines the response behavior of the household.

Given our definition of the household response propensity, we assume that

$$U_{R(h,t)} = \beta_0 + \beta_1 \, \mu_h + \beta_2 Y_{R(h,t-1)} + \varepsilon_{R\,h,t}, \quad \varepsilon_{R\,h,t} \sim N(0,1) \tag{4}$$

where the normally distributed error $\varepsilon_{R\,h,t} \sim N(0,1)$ has mean 0 and variance 1. Again, for model identification reasons, we can set the variance of the error to 1 without loss of generality. Intuitively, this specification amounts to the assumption that the deterministic part of $U_{R(h,t)}$ is a linear function of the household's propensity to respond $\tau_{h,t}$ defined in Eq. (1).

Because the error in this expression is normally distributed, the model for the household purchase variable is a probit model, conditional upon the long- and short-run elements of the propensity to respond construct. It should be noted that this model is only applied to households who are targeted by the firm. Households who do not receive a product offer cannot buy the product. For these households, $Y_{R(h,t)}$ must be equal to 0. Stated differently, we can only estimate the response model over the set of households at a particular time point who receive a product offer from the firm.

Properties of the Errors

To complete the specification of the model, we make two key assumptions about the errors in the firm targeting equation and the household response equation. First, we assume that these errors are mutually independent at each time point. Second, we assume that these errors are independent over time.

The first assumption amounts to the notion of conditional independence. The intuition is that the household's propensity to respond drives both firm behavior and household behavior. Consequently, conditional on the values of μ_h and $Y_{R(h,t-1)}$, we can assume that the selection and response error terms in this model are independent. Since different values of μ_h and $Y_{R(h,t-1)}$ lead to different values of selection and response, our model implies a natural correlation between observed selection and observed response across the household population. In other words, conditional independence allows for a simpler representation of the choice process without sacrificing the reality that selection and response are correlated. Conditional independence is a key element of model construction both in psychometrics (Lord & Novick, 1968; Fischer & Molenarr, 1995; Langeheine & Rost, 1988) and marketing science (e.g., Kamakura & Russell, 1989; Rossi et al., 1996).

The second assumption is necessary to prevent endogeneity issues from entering the model through the lagged response variable. Lagged response $Y_{R(h,t-1)}$ is already modeled in the system by the selection and response equations at time $t-1$. Given the value of propensity to respond at the previous period, the probability that $Y_{R(h,t-1)}$ equals one is a product of the probability of selection and the probability of response in period $t-1$. Consequently, in the context of the response model, the observed lagged response is only correlated with the error terms in previous time periods (i.e., periods $t-2$, $t-3$, $t-4$). Since the error terms are assumed independent over time, the lagged response $Y_{R(h,t-1)}$ cannot be correlated with the error

terms in the current period ($\varepsilon_{S\,h,t}$ or $\varepsilon_{R\,h,t}$). Thus, the inclusion of $Y_{R(h,t-1)}$ in the model does not create endogeneity problems.

It is important to notice that the model developed here does not suffer from the problems of selection bias and endogeneity noted in our earlier discussion of RFM models. Both μ_h and $Y_{R(h,t-1)}$ are independent of the errors in the selection and response models (Eqs. (3) and (4)), thus eliminating endogeneity from the specification. Moreover, as explained by Heckman (1979), biases in parameter estimation due to selection bias are entirely due to a nonzero correlation between the errors of the selection and response equations. Because the errors $\varepsilon_{S\,h,t}$ and $\varepsilon_{R\,h,t}$ are contemporaneously independent, selection bias is not present in the estimates generated from our model.

Model Estimation

The proposed model (Eqs. (2)–(4)) is a two-equation probit system with an underlying latent variable measuring the response propensity of each household. (The definition of the propensity to respond construct (Eq. (1)) is used to motivate the structure of the model, but the δ_1 and δ_2 coefficients are not explicitly estimated by our algorithm.) We calibrate the model by formulating the estimation problem using Hierarchical Bayes concepts and employing Markov Chain Monte Carlo (MCMC) technology to simulate draws from the posterior distribution of parameters (Gelman, Carlin, Stern, & Rubin, 1996). Details on the algorithm are presented in the appendix.

The convergence of the MCMC algorithm was checked using a procedure developed by Geweke (2001). In Geweke's approach, a second simulation, which uses a different (nonstandard) logic to draw the simulated values, is conducted following the initial MCMC analysis of the data. Because Geweke (2001) proves that the initial and the new simulation constitute Markov chains with the same stationary point, the researcher is able to check convergence by verifying that the posterior means and variances from the two simulations are the same. The results reported in this chapter passed this stringent convergence test.

APPLICATION

To understand the properties of the propensity to respond model, we use a customer database from a nonprofit organization. Our intention here is to

contrast the latent trait approach to a predictive model based solely on RFM. It is important to understand that the propensity to respond model does not make any use of traditional RFM variables. This difference is important because it allows us to compare the industry standard RFM approach to a formulation that ignores RFM variables. From a substantive point of view, this application is also designed to show that the propensity to respond model yields insights into the response characteristics of different types of mail solicitations used by the firm and the operating characteristics of the firm's current household selection policy.

Data Description

The data consist of the transaction records of a nonprofit organization that uses direct mail to solicit contributions from donors. Data are taken from the period October 1986 through June 1995. There is one record per past donor. Each record contains information on donor identification, postal code, donation history, and solicitation dates. Since the contribution codes and solicitation codes match, each contribution can be traced to a specific solicitation type and date.

We selected a random sample of 1,065 households for our analysis. Since we need a start-up time period to define RFM values, the final calibration sample contains 20 solicitations during the period from July 1991 to October 1994. The holdout sample for verification of the results contains four solicitations potentially available to these households during the period from November 1994 to March 1995. Overall, households receive mailings from as few as two times to as many as 11 times across 20 time periods.

A preliminary analysis of household donation behavior showed that the amount of money donated by household varies little over time. For example, if a given household donates \$5 on one occasion, the household is very likely to donate \$5 on every donation occasion. This fact allows us to regard the amount of the donation as a stable characteristic of the household, and concentrate only on the probability that the household decides to make a donation. Thus, the use of our model – a model that focuses only on the incidence of selection and response ($Y_{S(h,t)}$ and $Y_{R(h,t)}$) – is entirely appropriate for this application.

There are four major solicitation types, types A, B, C, and a miscellaneous type. Type A is the major solicitation type that shows the most frequent and regular mailings every three to six months. Type B includes the holiday mailings of December and January. Types C and miscellaneous are less

frequent than types A and B and do not show a regular pattern of mailing. We record these solicitations as types A and non-A. In some cases, the type of the solicitation sent to household is not recorded in the dataset. These unknown types are called "type unknown." By separating out the "type unknown" solicitations, we are able to study the characteristics of types A and non-A without making unwarranted assumptions. As we show subsequently, the selection and response characteristics of types A and non-A solicitations are decidedly different.

A postal code dataset is used to obtain a demographic description of each household. This dataset includes postal code, income index, percentage of households occupied by white, black, and Hispanic persons, percentage of households with one or more children under 18, persons per household, household median age, and median years of school for people aged 25 or more. Including gender information taken from the donation record, there are a total of nine demographic features. To improve the convergence of our estimation algorithm, we used principal components analysis to create a set of nine uncorrelated demographic variables. All nine principal component variables are used in the analysis.

Latent Trait Models

In our analysis of the donation dataset, we consider two variants of the propensity to respond model in this application. The most general model, called the "Dynamic model," takes the general form

$$U_{S(h,t)} = \alpha_{0k} + \alpha_{1k}\,\mu_h + \alpha_{2k}\,Y_{R(h,t-1)} + \varepsilon_{S\,h,t}, \quad \varepsilon_{S\,h,t} \sim N(0,1) \tag{5}$$

$$U_{R(h,t)} = \beta_{0k} + \beta_{1k}\,\mu_h + \beta_{2k}\,Y_{R(h,t-1)} + \varepsilon_{S\,h,t}, \quad \varepsilon_{S\,h,t} \sim N(0,1) \tag{6}$$

where k denotes type of solicitation (types A, non-A, or unknown), and the errors are mutually independent (contemporaneously and for all possible leads and lags). This is the model discussed earlier.

We also estimate a restricted model, called the "Long-Run model," in which the coefficients on lagged choice (α_{2k} and β_{2k} for all solicitation types k) are set to zero. This second model has the form

$$U_{S(h,t)} = \alpha_{0k} + \alpha_{1k}\,\mu_h + \varepsilon_{S\,h,t}, \quad \varepsilon_{S\,h,t} \sim N(0,1) \tag{7}$$

$$U_{R(h,t)} = \beta_{0k} + \beta_{1k}\,\mu_h + \varepsilon_{R\,h,t}, \quad \varepsilon_{R\,h,t} \sim N(0,1) \tag{8}$$

Note, in particular, that the implied selection and response probabilities in Eqs. (7) and (8) vary across households h, but *do not* vary over time t. This model is useful for two reasons. First, it allows us to judge whether the flexibility provided by Dynamic model leads to better forecasting performance. Second, it serves as a contrast to the RFM approach, which implicitly assumes that a household's response propensity varies continuously over time.

Traditional RFM Model

To benchmark the latent trait model, we consider a standard RFM probit model. This uncorrected RFM probit is the model specification, which is typically used by industry consultants.

The definitions of the RFM variables used in this research follow standard industry practice (David Sheppard Associates, Inc., 1999; Hughes, 1996). Recency is defined as the number of days since the last donation received by the firm. Frequency is defined as the total number of contributions made in the past up to the solicitation date. Monetary value is defined as the cumulative amount of contributions (in dollars) that the household spent previous to the solicitation date. Since the correlation between the frequency and monetary variables in our data is 0.95, only the recency and frequency variables are used in the RFM model. For this reason, model coefficients for frequency must be understood to incorporate the impact of monetary value as well.

The traditional RFM model is a binary probit system that forecasts the probability of household response. Formally, we write the utility of household response as

$$U_{R(h,t)} = \beta_{0k} + \beta_{1k}[\mathrm{Rec}]_{ht} + \beta_{2k}[\mathrm{Freq}]_{ht} + \varepsilon_{R\,h,t} \tag{9}$$

where the error $\varepsilon_{R\,h,t}$ has a normal distribution with means equal to 0, variance equal to 1. The subscript k in this model denotes the type of solicitation sent to the household: types A, non-A, and unknown. Here, Rec denotes recency and Freq denotes frequency. To connect this model with the observables in the donation dataset, we assume the probability that $U_{R(h,t)} > 0$ is the probability that binary variable $Y_{R(h,t)} = 1$ (household makes donation).

Assessing Forecast Accuracy

We assess model performance by predicting response in the holdout dataset. Since the response behavior is not observed when a household is not

selected, prediction is based on the response of the selected households only. To ensure comparability across models, we use the mean absolute deviation (MAD) statistic and the overall accuracy (hit rate for both purchases and nonpurchases). To take account of estimation uncertainty, these measures are computed using 2,000 simulated draws of parameters from the posterior distribution for the latent trait models and the RFM model estimated by MCMC. The mean of the 2,000 prediction measures are reported for all models. This procedure enables the forecast measures to be tested for the statistical difference.

The holdout prediction statistics of the two latent trait models and the RFM model are presented in Table 1. Overall, the latent trait models show performance better than the traditional RFM probit model in both MAD and the Overall Hit. The MAD of the Dynamic and Long-Run models is not statistically different, but the Overall Hit of the Dynamic model is statistically higher than the Long-Run model.

Latent Trait Model Coefficients

The pattern of coefficients for Dynamic latent trait model (Table 2) tells an interesting story. Beginning with demographic effects, it is important to note that six of the nine principal component variables capturing demographic effects are statistically insignificant. This indicates that demographics are generally a poor guide to the long-run propensity to respond μ_h trait of households. In line with the extensive marketing science literature on consumer heterogeneity (for a review, see Allenby and Rossi, 1999), most of the differences in long-run household buying behavior are unrelated to the set of demographics available for analysis.

Table 1. Accuracy of Holdout Data Forecasts.

	MAD (%)	Overall Hit (%)
Dynamic Model	31.30	78.71
Long-Run Model	31.34	78.57
RFM Probit Model	32.84	75.76

Notes: MAD of the Latent Trait Models are significantly lower than the RFM Probit Model ($p < .01$). MAD of the two Latent Trait Models are not significantly different ($p < .01$). Overall Hit of the Latent Trait Models are significantly higher than the RFM Probit Model ($p < .01$). Overall Hit of the Dynamic Model is significantly higher than the Long-Run Model ($p < .01$).

Table 2. Dynamic Propensity to Respond Model.

Demographics	Coefficient	Standard Deviation
Non-White	0.0192	0.0255
High income/education	−0.0679*	0.0301
White, children under 18	0.1336*	0.0328
Black male, non-Hispanic	−0.0957*	0.0399
Hispanic male	−0.0184	0.0421
Household size, age	−0.0366	0.0492
Children under 18, small household	−0.1061	0.0735
Low education/income	0.0892	0.1029
White and Black (non-ethnic)	−0.0423	0.1989

	Selection	Standard Deviation	Response	Standard Deviation
Type A Solicitation				
Intercept	−0.6731*	0.0170	−0.2699*	0.0379
Long-Run Propensity to Respond	0.1077*	0.0212	0.7278*	0.0609
Lagged Response	−0.5159*	0.0523	0.1175	0.1424
Type Non-A Solicitation				
Intercept	−0.5125*	0.0140	−1.0307*	0.0321
Long-Run Propensity to Respond	−0.0340*	0.0149	0.3618*	0.0533
Lagged Response	0.3959*	0.0489	−0.2449*	0.0989
Unknown Solicitation Type				
Intercept	0.1087*	0.0195	−0.2051*	0.0345
Long-Run Propensity to Respond	0.1750*	0.0352	0.8583*	0.0556
Lagged Response	1.3628*	0.2328	−0.1429	0.1877

Notes: Demographics are principal component variables derived from a postal code demographic dataset. Standard deviation indicates posterior standard deviation of the corresponding coefficient distribution. Parameter estimates denoted by an asterisk (*) are more than two standard deviations away from zero.

The pattern of results among the solicitation types is quite informative. Consider the differences between solicitation type A (the routine mailings) and solicitation type non-A (special mailings, often seasonal). (We do not discuss the type unknown solicitations because they are some unknown mixture of all solicitation types.) Note that the selection rules are quite different. Type A is sent to households who are better long-run prospects, but who have not donated recently. In contrast, type non-A mailings basically ignore long-run response, instead emphasizing households that have donated recently. Turning

to the response coefficients, it is clear that type A solicitations generate a much stronger long-run response than type non-A solicitations. In contrast, the significantly negative coefficient on lagged response for type non-A solicitations indicates that households who recently donated are unlikely to donate again.

Graphs of the selection and response curves, shown in Fig. 1, reinforce these general points. The horizontal axis, displaying the long-run propensity to respond trait, is restricted to the range of μ_h values characterizing the households in our dataset. The overall impression conveyed by Fig. 1 is that

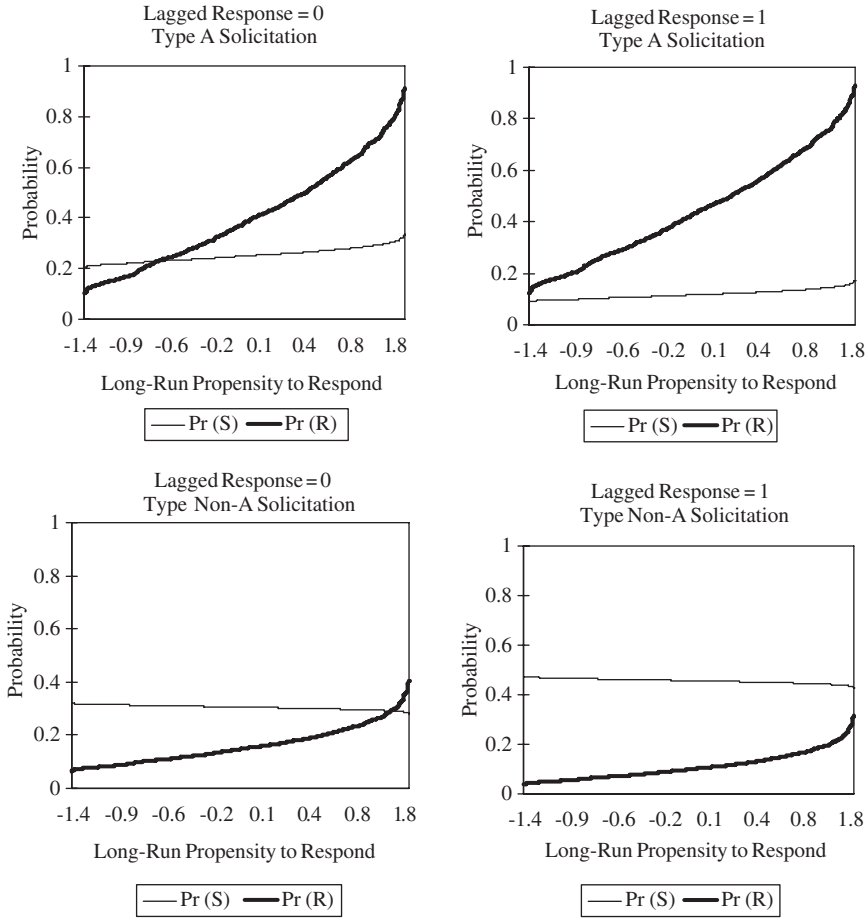

Fig. 1. Selection and Response Functions for Donation Dataset.

the decision to donate to this charitable organization is closely tied to the value of the propensity to respond measure. This, of course, is how the model is constructed. However, Fig. 1 also shows that the decision to mail a solicitation to a household depends very weakly on this trait. Managers, instead, seem to base mailing policy primarily on whether the household has donated in the recent past.

Substantively, these results suggest that type A solicitations work better for this charity because managers wait for a period of time after a donation before sending out a new solicitation and attempt to target households with long-run interests in the charity. In contrast, the type non-A solicitations are mailed out in a more opportunistic fashion, relying more on short-run response. It is possible that the firm views type A solicitations more in terms of household retention, and the type non-A solicitations more in terms of household acquisition. Nevertheless, given the response pattern for the type non-A solicitations, the current mailing policy for type non-A (mailing to those who have recently responded) is clearly counterproductive. This observation, along with the fact that the mailing policy of the routine type A solicitations is weakly linked to the household propensity trait, strongly argues that the firm's mailing policy could be improved by using the estimated propensity to respond as a guideline for mailing.

Summary

RFM variables are best regarded as behavioral indicators of an underlying interest in the firm's product or service. In this case, the product is the cause of the charitable organization. Although RFM variables provide some information on a household's propensity to respond trait, RFM variables are also impacted by the mailing policy selected by managers. The latent trait approach is superior in the sense that it separates the household trait (the decision to respond to a solicitation) from behavioral responses to the trait (the manager's decision to contact a household). Moreover, this trait is a stable characteristic of the household which cannot be affected by the firm.

CONCLUSION

This study develops a general procedure for estimating the response probabilities of households in database marketing. The proposed approach, based on a simultaneous selection–response formulation, assumes that each

household has a latent propensity to respond that impacts both the firm's decision to mail a solicitation and the household's decision to respond to a solicitation. Inclusion of the selection decision of the firm in the model recognizes the potential for selection bias; the propensity to respond construct solves the problem of endogeneity of RFM. Our empirical analysis showed that the Dynamic model yielded the best forecasting results. The latent trait model generates exogenous measures of the long-run propensity to respond to each household and provides the researcher with a tool to understand the effectiveness of current household solicitation policy.

Contributions of Research

Although recency, frequency, and monetary value are intuitively reasonable ways of measuring the attractiveness of households, constructing a predictive model using RFM variables is problematic. The underlying problem is that the RFM variables in a database are functions of both the household's interest in the product category and the firm's mailing policy. That is, a household's RFM profile depends on characteristics of the both the household and the firm. From a statistical point of view, this confounding of household behavior and firm decision behavior is particularly worrisome.

The proposed model, by explicitly considering the rule of household selection rules in generating the dataset, calibrates a household response model that can be generalized to future datasets. This generalizability is due to the fact that our model is free from the selection bias and RFM endogeneity problems that affect most conventional methodologies. Using the response coefficients from the model output along with knowledge of the household trait value and lagged response behavior $Y_{R(h,t-1)}$, a researcher can predict response behavior in a future scenario in which the firm's decision rules have changed. A major strength of the model is its ability to recover the true response characteristics of households from a customer database, even when the firm has attempted to optimize the mailing of solicitations. In principle, this is not possible with the RFM approach because the observed RFM profile depends on both past purchase behavior and whether or not managers selected the household for mailings.

Limitations and Extensions

Our model has several limitations, all of which provide avenues for future research. The current model predicts only the incidence of selection and

response. The most obvious extension is the construction of a model that predicts both probability of response and the dollar donation (or expenditure). This could be accomplished by changing the response equation to a Tobit model (Davidson & MacKinnon, 1993), in which the latent propensity to respond drives both incidence and dollar amount. The current model is also limited to the prediction of one response per household. Clearly, most catalog retailers sell a large array of products. These retailers often develop specialty catalogs emphasizing subsets of the product line, and target these catalogs to various segments in the customer database. By constructing a set of correlated latent response variables for different product subsets, the model could be further generalized to consider the basket of purchases made by a household. Taken together, these generalizations would permit the analyst to develop a global choice model for use by a multiple-category catalog retailer.

ACKNOWLEDGMENT

The authors thank the Direct Marketing Educational Foundation for providing access to the data used in this study. The authors also thank Professor John Geweke of the Department of Economics, University of Iowa, for many helpful suggestions on model specification and estimation.

REFERENCES

Allenby, G. M., & Rossi, P. E. (1999). Marketing models of consumer heterogeneity. *Journal of Econometrics*, *89*, 57–78.

Balasubramanian, S. K., & Kamakura, W. A. (1989). Measuring consumer attitudes toward the marketplace with tailored interviews. *Journal of Marketing Research*, *26*(August), 311–326.

Bawa, K., Srinivasan, S. S., & Srivastava, R. K. (1997). Coupon attractiveness and coupon proneness: A framework for modeling coupon redemption. *Journal of Marketing Research*, *34*(November), 517–525.

Berger, P. D., & Nasr, N. (1998). Customer lifetime value: Marketing models and applications. *Journal of Interactive Marketing*, *12*(Winter), 17–30.

Berry, M. J., & Linoff, G. S. (2000). *Mastering data mining: The art and science of household relationship management*. New York: Wiley.

Bitran, G. R., & Mondschein, S. V. (1996). Mailing decisions in the catalog sales industry. *Management Science*, *42*(9), 1364–1381.

Bult, J. R., & Wansbeek, T. (1995). Optimal selection for direct mail. *Marketing Science*, *14*(4), 378–394.

David Sheppard Associates, Inc. (1999). *The new direct marketing: How to implement a profit-driven database marketing strategy*. Boston: McGraw-Hill.

Davidson, R., & MacKinnon, J. G. (1993). *Estimation and inference in econometrics*. New York: Oxford University Press.

Fischer, G. H., & Molenarr, I. W. (1995). *Rasch models: Foundations, recent developments and applications*. New York: Springer.

Gelfand, A. E., & Smith, A. F. M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association*, *85*(2), 398–409.

Gelman, A., Carlin, J., Stern, H., & Rubin, D. (1996). *Bayesian data analysis*. New York: Chapman and Hall.

Geweke, J. F. (2001). *Getting it right: Checking for errors in likelihood based inference*. Working Paper. Department of Economics, University of Iowa.

Geweke, J. F. (2003). *Contemporary Bayesian econometrics and statistics*. Working Monograph. Department of Economics, University of Iowa.

Gonul, F., & Shi, M. Z. (1998). Optimal mailing of catalogs: A new methodology using estimable structural dynamic programming models. *Management Science*, *44*(9), 1249–1262.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, *47*(1), 153–161.

Hughes, A. M. (1996). Boosting response with RFM. *Marketing Tools*, *5*, 4–10.

Jeuland, A. P. (1979). Brand choice inertia as one aspect of the notion of brand loyalty. *Management Science*, *25*(July), 671–682.

Jonker, J. J., Paap, R., & Frances, P. H. (2000). *Modeling charity donations: Target selection, response time and gift size*. Economic Institute Report 2000-07/A. Erasmus University Rotterdam.

Kamakura, W. A., Ramaswami, S. N., & Srivastava, R. K. (1991). Applying latent trait analysis in the evaluation of prospects for cross-selling of financial services. *International Journal of Research in Marketing*, *8*(4), 329–349.

Kamakura, W. A., & Russell, G. J. (1989). A probabilistic choice model for market segmentation and elasticity structure. *Journal of Marketing Research*, *26*(November), 271–390.

Langeheine, R., & Rost, J. (1988). *Latent trait and latent class models*. New York: Plenum.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA: Addison-Wesley.

Robert, C. P., & Casella, G. (1999). *Monte Carlo statistical methods*. New York: Springer-Verlag.

Rossi, P. E., McCulloch, R., & Allenby, G. (1996). The value of household information in target marketing. *Marketing Science*, *15*(Summer), 321–340.

Schmid, J., & Weber, A. (1997). *Desktop database marketing*. Chicago, IL: NTC Business Books.

Seetharaman, P. B., & Chintagunta, P. K. (1998). A model of inertia and variety-seeking with marketing variables. *International Journal of Research in Marketing*, *15*(1), 1–17.

Winer, R. S. (2001). A framework for customer relationship management. *California Management Review*, *43*(Summer), 89–105.

APPENDIX. ESTIMATION OF DYNAMIC PROPENSITY TO RESPOND MODEL

The Dynamic Propensity to Respond model is formulated in a Hierarchical Bayesian fashion (Gelman et al., 1996) and estimated using MCMC procedures (Robert & Casella, 1999). Here, we sketch the procedure used to estimate this model. The Long-Run Propensity to Respond model is estimated in a similar fashion by simply deleting the lagged response variable from both selection and response equations.

Dynamic Response Model

The Dynamic Propensity to Respond model is formulated in the following manner. The household-specific long-run propensity to respond for each household h is considered to be an independent, random draw from the normal distribution

$$\mu_h \sim N(X_h\gamma, 1.0) \tag{A.1}$$

where μ_h is the long-run propensity to respond construct, X_h a vector of household demographics, and γ a vector of parameters. We assume, without loss of generality, that the precision of the normal distribution (inverse of the variance) is equal to one.

At each time point t, the household has two potential observations: a binary variable $Y_{S(h,t)}$ reporting whether household h was sent a solicitation ($=1$) or not ($=0$); and a binary variable $Y_{R(h,t)}$ reporting whether household h made a donation ($=1$) or not ($=0$). The selection and response models are formulated as two independent utility models, conditional on the long-run propensity to respond construct μ_h and on whether the household made a donation during the last solicitation $Y_{R(h,t-1)}$. Formally, we write the utility of selection as

$$U_{S(h,t)} = \alpha_0 + \alpha_1\mu_h + \alpha_2 Y_{R(h,t-1)} + \varepsilon_{S\,h,t} \quad \varepsilon_{S\,h,t} \sim N(0,1) \tag{A.2}$$

and the utility of response as

$$U_{R(h,t)} = \beta_0 + \beta_1\,\mu_h + \beta_2 Y_{R(h,t-1)} + \varepsilon_{R\,h,t} \quad \varepsilon_{R\,h,t} \sim N(0,1) \tag{A.3}$$

for all time point t of each household h. The two errors $\varepsilon_{S\,h,t}$ and $\varepsilon_{R\,h,t}$ are assumed independent both at time t, and over all possible pairs of past and future time points. In our empirical work, we allow the parameters of (A.2)

and (A.3) to depend on the type of solicitation sent by the firm. However, to simplify the exposition here, we ignore this feature of the model in the equations later.

We observe a mailing to a household ($Y_{S(h,t)} = 1$) when the utility of selection is greater than zero, and no mailing ($Y_{S(h,t)} = 0$) when the utility of selection is less than or equal to zero. In a similar fashion, we observe a donation ($Y_{R(h,t)} = 1$) when the utility of response is greater than zero, and no donation ($Y_{R(h,t)} = 0$) when the utility of response is less than or equal to zero. Eqs. (A.2) and (A.3) form a two-equation binary probit system in which selection and response variables are independent, conditional on the values of μ_h and $Y_{R(h,t-1)}$.

Note that when $Y_{S(h,t)} = 0$ (no mailing to the household), then we must observe that $Y_{R(h,t)} = 0$ (no donation is made). That is, when the household is not sent a mailing, we observe no response, but do not know whether or not the household would have responded if given the opportunity. For this reason, it is necessary to regard the donation response as missing whenever $Y_{S(h,t)} = 0$. Accordingly, in the development below, it is understood that Eq. (A.3) is dropped from the model for all combinations of h and t for which $Y_{S(h,t)} = 0$.

Prior Distributions

The prior distributions for γ, α, and β are assumed to be normal. Diffuse priors are chosen to allow the observed data to dominate the analysis. Specifically, we assume that $\gamma \sim N[0, \cdot 025\ I(d)]$, $\alpha \sim N[0, \cdot 025\ I(3)]$, and $\beta \sim N[0, \cdot 025\ I(3)]$ where d is the number of demographic variables and $I(z)$ denotes a (square) identity matrix of dimension z. Note that we are using Bayesian convention of writing a normal distribution as $N(m, p)$ where m is the mean and p the precision (the inverse of the variance).

Full Conditional Distributions

After constructing the posterior distribution for the model, we derive the full conditional distributions of the parameters. This leads to the following relations:

$$f[\mu|\text{else}] \sim N[\text{mean}(\mu), \text{prec}(\mu)] \tag{A.4}$$

where $\text{mean}(\mu) = \text{prec}(\mu)^{-1}\{X\gamma + \Sigma_t(\alpha_1 U_S) - \alpha_0\alpha_1 T - \Sigma_t(\alpha_1\alpha_2 Y) + \Sigma_t(\beta_1 U_R) - \beta_0\beta_1 T - \Sigma_t(\beta_1\beta_2 Y)\}$ and $\text{prec}(\mu) = (1 + \alpha_1^2 T + \beta_1^2 T)^* I(H)$. Here, $\text{mean}(\cdot)$ and $\text{prec}(\cdot)$ are the mean and precision of a normal distribution, X is $(H \times d)$ matrix of demographics, U_S and U_R are $(H \times T)$ matrices of utility of selection and utility of response, respectively, Y is $(H \times T)$ matrix of lagged response, and H the number of households, and T the number of time periods.

$$f[\gamma|\text{else}] \sim N[\text{mean}(\gamma), \text{prec}(\gamma)] \tag{A.5}$$

where $\text{mean}(\gamma) = \text{prec}(\gamma)^{-1}\{X'\mu + 0_{(d\times 1)}(.025)\}$, $\text{prec}(\gamma) = X'X + (.025)\ I(d)$, and $0_{(d\times 1)}$ is $(d \times 1)$ vector of zeros.

$$f[U_S|\text{else}] \sim \text{truncated } N[\bar{U}_S, I(H^*T)] \tag{A.6}$$

where the elements of $(H \times T)$ matrix of $\bar{U}_S$ are obtained from the deterministic elements on the right-hand side of Eq. (A.2). The notation truncated $N(a,b)$ denotes a truncated normal with upper bound $= 0$ if $Y_{S(h,t)} = 0$ and lower bound $= 0$ if $Y_{S(h,t)} = 1$.

$$f[U_R|\text{else}] \sim \text{truncated } N[\bar{U}_R, I(H^*T)] \tag{A.7}$$

where the elements of $(H \times T)$ matrix of $\bar{U}_R$ are obtained from the deterministic elements on the right-hand side of Eq. (A.3). The notation truncated $N(a,b)$ denotes a truncated normal with upper bound $= 0$ if $Y_{R(h,t)} = 0$ and lower bound $= 0$ if $Y_{R(h,t)} = 1$.

$$f[\alpha|\text{else}] \sim N[\text{mean}(\alpha),\ \text{prec}(\alpha)] \tag{A.8}$$

where $\text{mean}(\alpha) = \text{prec}(\alpha)^{-1}\{\mu Y^{*\prime} U_S^* + 0_{(3\times 1)}(.025)\}$ and $\text{prec}(\alpha) = \mu Y^{*\prime}\mu Y^* + (.025)I(3)$. Here, μY^* is $(H^*T \times 3)$ matrix that contains a $(H^*T \times 1)$ vector of ones, a vector of $\mu(H \otimes T)$, and a $(H^*T \times 1)$ vector of $Y_{R(h,t)}$. U_S^* is U_S $(H \otimes T)$ vector. $0_{(3\times 1)}$ is (3×1) vector of zeroes.

$$f[\beta|\text{else}] \sim N[\text{mean}(\beta),\ \text{prec}(\beta)] \tag{A.9}$$

where $\text{mean}(\beta) = \text{prec}(\beta)^{-1}\ \{\mu Y^{*\prime} U_R^* + 0_{(3\times 1)}\ (.025)\}$ and $\text{prec}(\beta) = \mu Y^{*\prime}\ \mu Y^* + (.025)I(3)$.

Here, μY^* is $(H^*T \times 3)$ matrix that contains a $(H^*T \times 1)$ vector of ones, a vector of $\mu(H \otimes T)$ and a $(H^*T \times 1)$ vector of $Y_{R(h,t-1)}$. U_R^* is U_R $(H \otimes T)$ vector. $0_{(3\times 1)}$ is (3×1) vector of zeros.

Markov Chain Monte Carlo Algorithm

To estimate the parameters of the model, we use the Gibbs Sampler (Gelfand & Smith, 1990). Starting with the vector of long-run propensity to respond parameters μ in Eq. (A.4), we successively sample the parameters from each equation in turn (i.e., sample the parameters of Eqs. (A.4)–(A.9) in order, then repeat the sequence). The stationary point of this Markov Chain contains the model parameters (Gelman et al., 1996).

For each of the models, we ran a chain of 20,000 simulates. We used the last 5,000 simulates to compute posterior means and variances. The convergence of this algorithm was checked using a procedure developed by Geweke (2001, 2003). In this approach, a second simulation, which uses a different (nonstandard) logic to draw the simulated values, is conducted. Because the underlying theory indicates that the initial and the new simulation constitute Markov chains with the same stationary point, the researcher can check convergence by verifying that the posterior means and variances of the two simulations are the same. The results reported in this chapter passed this stringent convergence test.

This page intentionally left blank

PART III
FORECASTING METHODS AND EVALUATION

This page intentionally left blank

A NEW BASIS FOR MEASURING AND EVALUATING FORECASTING MODELS

Frenck Waage

ABSTRACT

Assume that we generate forecasts from a model $y = cx + d + \xi$. *The constants "*c*" and "*d*" are placement parameters estimated from observations on* x *and* y, *and* ξ *is the residual error variable.*

Our objective is to develop a method for accurately measuring and evaluating the risk profile of a forecasted variable y. *To do so, it is necessary to first obtain an accurate representation of the histogram of a forecasting model's residual errors. That is not always so easy because the histogram of the residual* ξ *may be symmetric, or it may be skewed to either the left of or to the right of its mode. We introduce the probability density function (PDF) family of functions because it is versatile enough to fit any residual's locus be it skewed to the left, symmetric about the mean, or skewed to the right. When we have measured the residual's density, we show how to correctly calculate the risk profile of the forecasted variable* y *from the density of the residual using the PPD function. We achieve the desired and accurate risk profile for* y *that we seek. We conclude the chapter by discussing how a universally followed paradigm leads to misstating the risk profile and to wrongheaded decisions by too freely using the symmetric Gauss–normal function instead of the PPD function. We expect that this chapter will open up many new avenues of progress for econometricians.*

Advances in Business and Management Forecasting, Volume 6, 135–155

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006009

1. INTRODUCTION

A forecast predicts the future consequences of present choices, and estimates the probability of each consequence occurring, which is the risk profile of the forecast. Mathematical forecasting models take many forms including, but not limited to, the following:

$$y = f(x) + \xi$$
$$y_t = f(y_{t-1}, y_{t-1}, y_{t-3}) + \xi$$
$$y_t = f(y_{t-1}) + f(x) + \xi$$

We shall, in this chapter, let the very general formulation of a forecasting model $y = f(x) + \xi$ represent any of the functions that might be used in practice. The variable y is the dependent random variable whose values are to be forecasted, the term $f(x)$ is an arbitrary function of x and also possibly of lagged y variables, x measures the driving forces behind y, and ξ measures the model's residual errors. Once a forecasting model has been finalized, its residual or forecasting errors ξ are calculated from $\xi = y - f(x)$. Mathematical forecasting functions are discussed by Makridakis and Wheelwright (1978), Granger (1980), Levenbach and Cleary (1981), Cleary and Levenbach (1982), Box and Cox (1964), Graybill (1961), Mosteller and Tukey (1977), Neter, Kutner, Wassermann, and Nachtsheim (1996), Render, Stair, and Hanna (2006), and Weisberg (1985).

To obtain the risk profile of the forecast, first calculate the probability density, or histogram, of the forecasting model's residual errors. Second, identify a mathematical function that accurately fits the observed residual density or its histogram. This mathematical function measures the probability density of the residuals. From the now known residual density function calculate the probability density of the forecasted variable y. The density of y is the risk profile of the forecasted variable y. There are two basic approaches to identify the mathematical function, which most accurately fits the histogram of the residuals. These two approaches are now discussed.

Approach 1: Fit every known PDF, one at a time to the observed residual histogram. This includes fitting all of the normal, the Poisson, the binomial, the Weibull, the beta, the gamma, and many more to the histogram. The address on Internet http://wikipedia.org/wiki/Probability-distribution presents an adequate listing. The advantage with this approach is that a mathematical function will be identified, which provides the best fit. The disadvantage is that a very considerable amount of time will be consumed finding it.

Approach 2: Identify a single family of functions, which alone can take all of the different loci that it takes many individual functions to reveal in Approach 1. This one family of functions too is capable of supplying the locus, which best fits the residuals' histogram. The advantage of fitting only one family of functions is that only the function's placement parameters need to be estimated. The time consumed in doing this may be short. The disadvantage includes that statistical tests and confidence statements may have to be developed for this new function, and this may be both time consuming and demanding.

The purpose of this chapter is to introduce the single family of functions alluded to by Approach 2. That function is a polynomial probability density that we shall name PPD. By varying its placement parameters it will alone take all of the different loci that Approach 1 needs many functions to identify. The integral defined by Eq. (4) is the fundamental relationship. Cross-multiplications on Eq. (4) generate Eq. (5). Eq. (5) is the PPD function we introduce and discuss in this chapter. It possesses singly the capabilities to take all the different loci, which Approach 1 can deliver only with the help from very many different functions, by simply varying its placement parameters.

This chapter

(1) develops methods for fitting the new PPD function to any residual's histogram. The fitted PPD is the probability density of the residuals ξ;
(2) calculates the PDF of the random forecasted variable y from the known probability density of the forecasting model's residuals ξ. This PDF of y is the risk profile of the forecasted variable y; and
(3) demonstrates effective ways of using the risk profile in decision making.

The first step in this program is to develop the residual's histogram. How to do this is discussed next.

2. MEASURING THE HISTOGRAM OF THE RESIDUALS ξ

The forecasts of y_t for time period t are generated from a model $y_t = f(x_t) + \xi_t$. The residual errors of model are calculated from $\xi_t = y_t - f(x_t)$. The histogram of the residual errors is calculated by ordering the values of ξ from ξ's lower bound L to its upper bound M. The interval from L to M is subdivided into cells of equal width. Each observed ξ value is placed in the

appropriate cell. The number of ξ values in each cell is counted. The cell count is converted into percentages of the total number of observed ξ values. The percentages sum to 100%. This defines the histogram, and it can now be visualized. Graph the cell percentages along the ordinate axis and the cell widths (or cell midpoints) along the abscissa. The resulting graph reveals the residual's histogram.

2.1. An Application: Measuring the Residuals' Histogram

The linear regression model (1) was created from the 104 observations on x and y that have been tabulated in Table A1 in Appendix.

$$\begin{array}{ccc} y_t = 0.24883 x_t & + & 4.11807 \\ (0.2216) & & (0.0036) \\ 18.58 & & 67.90 \end{array} \tag{1}$$

where values in parentheses represent standard errors and 18.58 and 67.90 represent t-values.

$$\xi_t = y_t - 0.24883\, x_t - 4.11807 \tag{2}$$

The regression statistics are

$$\begin{aligned} &r^2 = 0.978, \quad \text{standard error} = 1.121914, \quad \text{variance}(\xi) = 1.258691, \\ &F = 4610.8, \quad E(\xi) = -0.0135 \end{aligned} \tag{3}$$

Using judgment, we determined that $L_\xi = -4.0$ and $M_\xi = +2.4$. The residual therefore has a finite range defined by $-4.0 \le \xi \le 2.4$. Fig. 1 depicts the residual's histogram and density. That histogram is visually skewed to the left.

We now seek a mathematical function that will correctly represent the locus of this histogram.

3. FITTING A MATHEMATICAL FUNCTION TO THE HISTOGRAM

To identify the function, which fits any given histogram best, we shall apply Approach 2. We shall create a single family of functions, which is so versatile in its capability to take different loci that it provides the best fit to any given histogram. This function is fundamentally defined by the integral (4).

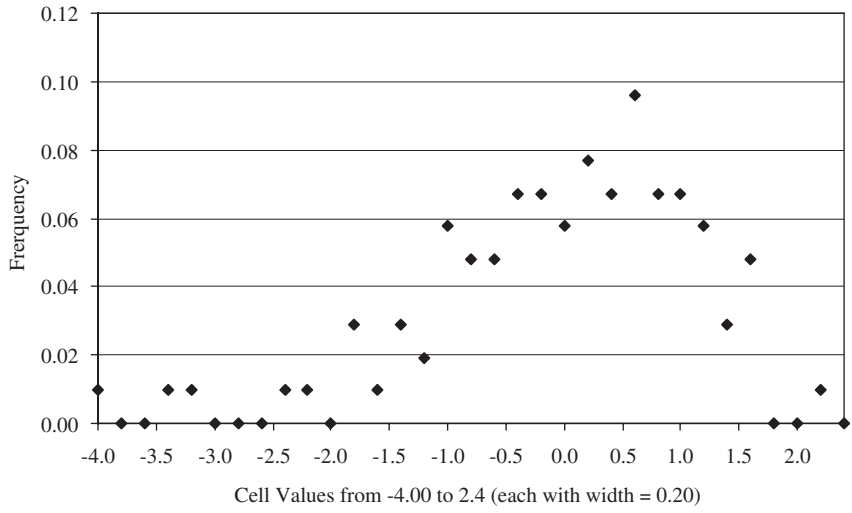

Fig. 1. The Histogram of the Residuals.

A continuous variable ξ varies in the range $(L_\xi \leq \xi \leq M_\xi)$, where L_ξ is the lower bound and M_ξ an upper bound on ξ. m and n measure the placement parameters. They are restricted to the values $m > -1$ and $n > -1$.

$$\int_{L_\varepsilon}^{M_\varepsilon}(\xi - L_\xi)^m(M_\xi - \xi)^n d\xi = (M_\xi - L_\xi)^{m+n+1} \quad \frac{\Gamma(m+1)\Gamma(n+1)}{\Gamma(m+n+2)} \tag{4}$$

The integral (4) is a generalization of the beta integral. Proofs of the integral are found in books on advanced calculus in sections that discuss gamma, beta, and related functions. Useful references are Allen and Unwin (1964), Davis (1959), Haible and Papanikolaou (1997), Havil (2003), Arfken and Weber (2000), and Abramowitz and Stegun (1972). Cross multiplication on Eq. (4) yields Eq. (5). Eq. (5) is the new function. It is the PPD family of functions, and it is a proper PDF.

$$f(\xi|m+1, n+1; L_\xi, M_\xi) = (M_\xi - L_\xi)^{-m-n-1} \times \frac{\Gamma(m+n+2)}{\Gamma(m+1)\Gamma(n+1)}(\xi - L_\xi)^m(M_\xi - \xi)^n \tag{5}$$

The locus of Eq. (5) will be skewed to the left of its mode if $m > n$;
The locus of Eq. (5) will be symmetric about its mean if $m = n$; and
The locus of Eq. (5) will be skewed to the right of its mode if $m < n$.

Along the abscissa, the values of ξ are concentrated on the interval $L_\xi \leq \xi \leq M_\xi$.

Integral (4) is not new. It has, however, not been used to any extent which I have been able to detect in regression modeling or in management science modeling. I consider Eq. (5) therefore to be new and unknown to these sciences. It is a significant function, because it can do most of the estimation work on symmetric forms that the normal function can do. And, it can do most of the estimation work on skew forms, which the normal function cannot do. The function should therefore attract considerable interest once it has been brought to light. This function is discussed by Lebedev (1972), Wilson (1912), and Hildebrand (1962). When the placement parameters m and n are known, the expected value of ξ, $E(\xi)$, and the variance of ξ, $V(\xi)$ of Eq. (5) are calculated from formulas (6) and (7). The formulas are calculated directly from Eq. (5) using the definitions of expectation and variance. Methods of calculating the moments are discussed by Green and Carroll (1978), Hair, Black, Babin, Anderson, and Tatham (2006), Harris (2001) and Wilks (1962).

$$E(\xi) = (M_\xi - L_\xi)\left(\frac{(m+1)}{(m+n+2)}\right) + L_\xi \tag{6}$$

$$V(\xi) = (M_\xi - L_\xi)^2 \quad \frac{(m+1)(n+1)}{(m+n+2)^2(m+n+3)} \tag{7}$$

We will now discuss two practical methods for fitting the PPD defined by Eq. (5) to any observed histogram: method A and method B. A vast literature addresses the problem of fitting mathematical functions to observations. Some of the useful references are: Brownlee (1965), Daniel and Wood (1999), Maddala (1977), Render, Stair, and Balakrishnan (2006), Kendall (1951), Kendall and Stuart (1958), Feller (1957), Feller (1966), Hines and Montgomery (1972), and Wilks (1962).

3.1. Using the PPB Function and Fitting Method A

The problem we confront is: Find the values of m and n in Eq. (5), which minimize the sum of squared differences between the locus of Eq. (5) when these values for m and n are used, and the locus of the observed histogram when L and M are known constants. This method requires that we solve

the non-linear program that finds the values for m and n, which minimize Eq. (8) while they satisfy Eqs. (9) and (10).

$$\min_{m,n} \Omega = \sum (F(\xi o) - F(\xi c))^2 \tag{8}$$

$$m \geq 0 \tag{9}$$

$$n \geq 0 \tag{10}$$

This non-linear program can be solved by the software SOLVER available in Microsoft's spreadsheet EXCEL. To set the problem up in EXCEL, so that the solution can be found, proceed as shown in the following spreadsheet. The spreadsheet addresses are by rows 1, 2, 3, etc. and by columns A, B, C, etc. The complete spreadsheet has 104 observations on the residuals, which Table A1 in appendix holds. Following is an abstract of the full spreadsheet:

Row	Column A	Column B	Column C	Column D	Column E
1	Value of m	Value of n	L	M	
2	4.8748	2.4890	− 4.00	+ 2.40	
4					
5	Residual cells width = 0.2	Count of residuals in each cell	Percentage of ξ calculated from column B for each cell	Percentage of ξ calculated from Eq. (5) at cell midpoints	Minimize this sum of the squared differences
			$F(\xi_o)$	$F(\xi_c)$	$(F(\xi_o) - F(\xi_c))^2$
6	− 4.00	**1**	0.00962	**0.0000**	0.0001
7	− 3.80	0	0.00000	**0.0000**	0.0000
8	− 3.60	0	0.00000	**0.0000**	0.0000
9	− 3.40	**1**	0.00962	**0.0000**	0.0001
10	− 3.20	1	0.00962	**0.0001**	0.0001
30	1.80	0	0.00000	**0.0201**	0.0004
31	2.00	0	0.00000	**0.0104**	0.0001
32	2.20	1	0.00962	**0.0035**	0.0000
33	2.40	0	0.00000	**0.0004**	0.0000
34	Column totals	**104**	**1.00000**	**1.0000**	**0.0031**

1. Complete, by calculating them, spreadsheet columns A–E.
2. Enter in row 1 the text shown in columns A–D. Enter the values shown in row 2 columns A–D. $L = -4.00$ and $M = +2.4$ are lower and upper bounds on the residual. They do not change. Enter in A2 and B2 the arbitrary starting values $m = 1.00$ and $n = 1.00$, which the solution algorithm SOLVER needs to start. The solution software SOLVER will use these in its first "round," and find better values in each subsequent round terminating the algorithm after many "rounds" on the optimal values for m and n. These optimal values are those that will minimize the sum of squared differences in column E, and those we seek. The optimal values of m and n will be placed in addresses A2 and B2 replacing the initial starting values.
3. In EXCEL, the software SOLVER will find the desired values of m and n. To find the optimal solutions for m and n using SOLVER in EXCEL, proceed as follows.
4. Complete all entries in the table shown above for all rows and all columns A–E.
5. In EXCEL, click on TOOLS. Under TOOLS, click on SOLVER or on PREMIUM SOLVER. A Window labeled SOLVER PARAMETERS opens on your screen. Type the following information into this window:

   - Where it says *Set Target Cell* type the addresses that hold the sum of the squared differences, which you intend to minimize. Above that is \$E\$34 (Column E, Row 34). This is your objective to be minimized
   - *Next*, click on the *MIN button* because you wish to minimize the sum of squares in \$E\$34.
   - Where it says *By Changing Cells* type the address where the values for m and n are. Above that is \$A\$2:\$B\$2. This is the address that holds the starter values for m and n. It is also where the optimal values for m and n will be placed by the software program.
   - Where it says *Subject to* constraints click on "*Add*" and type in \$A\$2 $\geq$ 0.00, then click on "*Add*" again and enter \$B\$2 $\geq$ 0.00, and then click on "OK".
   - *Click on the button OPTIONS*. Under OPTIONS click on ASSUME NON-NEGATIVE then on OK.
   - *Finally, to get the software to calculate the optimal values of* m *and* n*, click on the button SOLVE* in the upper right corner of the SOLVER PARAMETERS window. SOLVER calculates the sum of squares for each allowed pair (m, n). It finds the pair (m, n) that generates the smallest sum. It prints the optimal pair in the location A2 for m and B2 for n. These are the minimizing values for m and n, which you seek. The minimized sum of squared differences is keyed in address \$E\$34.

3.2. An Application of Method A: The PPD Function Fitted to the Histogram

The regression application first recorded in Eqs. (1), (2) and (3) is continued here. The residuals were calculated from Eq. (2). All the 104 residual values are shown in Table A1 in appendix. The foregoing spreadsheet shows an abstract. The entire spreadsheet columns A–E were completed for this application. The 104 observed residual observations were placed in their cells. Each cell is 0.20 wide. The two smallest cells are (-4.00 to -2.81), (-2.80 to -2.61). SOLVER was used to obtain the optimal solution when $L_\xi = -4.0$ and $M_\xi = 2.4$. The non-linear program solved by SOLVER calculated the optimal values of m and n for the PPD function to be

$$m = 4.8748, \quad n = 2.4890, \quad L = -4.00, \quad M = +2.40$$

Substituting these parameter values into the PPD function (5) produces the PPD function (10a)

$$\begin{aligned} f(\xi|m+1, n+1; L_\xi, M_\xi) &= f(\xi|4.8748+1, 2.4890+1; -4.00, +2.40) \\ &= (6.40)^{-8.3638} \frac{\Gamma(9.3638)}{\Gamma(5.8748)\Gamma(3.4890)} (\xi + 4.00)^{4.85} (2.40 - \xi)^{2.28} \end{aligned} \tag{10a}$$

The mean and the variance of the residuals of the function (10a) are obtained from the moments $E(\xi)$ and $V(\xi)$ of this function. The moments are calculated by substituting the optimal values for $m = 4.8748$ and $n = 2.4890$ into formulas (6) and (7). Calculations yield

$$E(\xi) = 0.0153, \quad V(\xi) = 0.9239$$

Fig. 2 graphs the residuals and the PPD function (10) which have been fitted to the residuals' histogram. This PPD density is skewed to the left of the mode because $m > n$. The mean $E(\xi)$ is the vertical dotted line in the graph. The minimum sum of the squared differences = 0.0031.

There is a second method for fitting the parameters m and n to observations, which is discussed next.

3.3. Fitting Method B

Solving Eqs. (6) and (7) simultaneously for m and n produce formulas (11) and (12). If we now knew $E(\xi)$ and $V(\xi)$, the upper bound M_ξ and the lower

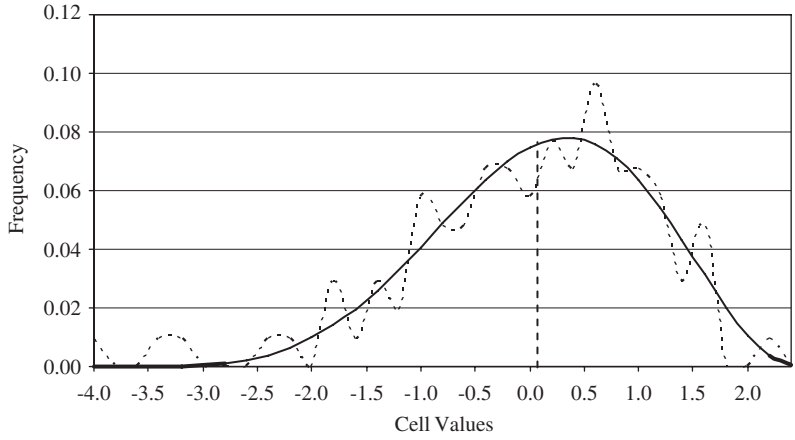

Fig. 2. The Residual Histogram (Continuous Dashed Curve). The Fitted PPD function (Continuous Bold Curve)

bound L_ξ of the residuals, we could calculate from formulas (11) and (12) the values of m and n, which are consistent with the observed $E(\xi)$ and $V(\xi)$, M_ξ and L_ξ . So calculate $E(\xi)$, $V(\xi)$, M_ξ, and L_ξ from the observed residuals. Substitute these into (11) and (12) to obtain corresponding placement parameter values for m and n.

$$m = \left(\frac{(E(\xi) - L_\xi)}{V(\xi)(M_\xi - L_\xi)}\right)\big((E(\xi) - L_\xi)(M_\xi - E(\xi)) - V(\xi) - 1.00\big) \qquad (11)$$

$$n = \left(\frac{(M_\xi - E(\xi))}{V(\xi)(M_\xi - L_\xi)}\right)\big((E(\xi) - L_\xi)(M_\xi - E(\xi)) - V(\xi) - 1.00\big) \qquad (12)$$

Using the values $E(\xi) = -0.0135$, $V(\xi) = 1.2548$, $L_\xi = -4.00$, and $M_\xi = 2.40$ in formulas (11) and (12), we calculate that $m = 3.15$ and $n = 1.51$. This too produces a "tight" fit, but method A has a smaller sum of the squared differences, and provides therefore the best estimate.

Fitting methods A and B give us the placement parameters m and n, which best fit the PPD function to the observed residuals histogram. This gives us the density of the residuals. It is, however, the density of the forecasted variable y we are interested in. The next section shows how the probability density of y is calculated from the known density of ξ.

4. CALCULATING THE DENSITY OF Y FROM THAT OF ξ

The variables y and ξ are linked by a model $y = f(x) + \xi$, or equivalently, by $\xi = y - ax - b$. To calculate the probability density of y from the known density of ξ, execute the following operations:

Step 1. Calculate the Jacobi transform $|J| = \partial\xi/\partial y$ from Eq. (2). The Jacobi transform $|J| = \partial\xi/\partial y = 1.00$. The Jacobi transform is discussed by Brownlee (1965), Yamane (1962), and Hines and Montgomery (1972).

Step 2. Substitute the residual function $\xi = y - ax - b$ given by Eq. (2) into the PPD function (5) in order to eliminate ξ from Eq. (5) and to introduce y into it. The polynomial term in Eq. (5) will become $(y - ax - b - L_\xi)^m (M_\xi - y + ax + b)^n$.

Step 3. Simplify this polynomial term by defining the lower bound of y to be $L_y \equiv ax + b + L_\xi$ and the upper bound on y $M_y \equiv ax + b + M_\xi$. Using these expressions for L_y and M_y in the polynomial terms, we obtain the simplified expression given by $(y - L_y)^m (M_y - y)^n$.

Step 4. Obtain the PDF of y by executing the operations in Eq. (13). Eq. (13) is the probability density of the forecasted variable Y, and it gives us the risk profile of the forecast y.

$$\begin{aligned} g(y)dy &= g(f(y))|J|dy = g(y|m+1, n+1; x, L_y, M_y)dy \\ &= (M_y - L_y)^{-m-n-1} \frac{\Gamma(m+n+2)}{\Gamma(m+1)\Gamma(n+1)} [(y - L_y)]^m [(M_y - y)]^n \end{aligned} \tag{13}$$

$M_y = ax + b + M_\xi$ is the upper bound on y, and
$L_y = ax + b + L_\xi$ is the lower bound on y.

For each possible "x" in the regression Eq. (1), there exists a density function that measures the probability of any y value occurring for that given x value. The kernel of Eq. (5) and of Eq. (13) has the same form. We can therefore use formulas (6) and (7) to calculate $E(y)$ and $V(y)$ for Eq. (13) using the known m and n, but replacing the bounds L_ξ and M_ξ in the formulas with the new bounds for y, L_y, and M_y. Formulas (11) and (12) calculate m and n using $E(y)$, $V(y)$, M_y, and L_y in lieu of $E(\xi)$, $V(\xi)$, M_ξ, and L_ξ. Eq. (12) is the desired risk profile of the forecasted values for y.

4.1. An Application: Calculating the Density of y from the Density of ξ

Using the results from the regression analysis from Eqs. (1), (2) and (3), we calculate the upper bound M_y and the lower bound L_y of y.

$$L_y = ax + b + L_\varepsilon = 0.24883x + 4.11807 - 4.00 = 0.24883x + 0.11807$$
$$M_y = ax + b + M_\varepsilon = 0.24883x + 4.11807 + 2.40 = 0.24883x + 6.51807$$

The continuous random forecasted variable y varies in the finite range bounded by

$$L_y = 0.11807 + 0.24883x \le y \le 6.51807 + 0.24883x = M_y$$

The optimal placement parameters remain unchanged as $m = 4.8748$, $n = 2.4890$.

The density function of y is skewed to the left of the mode because $m > n$. Use these values in Eq. (13) to obtain the PPD density, which governs y. It is given by Eq. (14).

$$\begin{aligned} g(y|m+1, n+1; x, L_y, M_y) &= g(y|4.8748 + 1, 2.4890 + 1; 0.11807 \\ &\quad + 0.24883x, 6.51807 + 0.24883x) \\ &= (6.4)^{-8.3638} \frac{\Gamma(9.3638)}{\Gamma(5.8748)\Gamma(3.4890)} \\ &\quad \times [(y - 0.118 - 0.2488x)]^{4.8748} \\ &\quad \times [(6.518 + 0.2488x - y)]^{2.489} \end{aligned} \tag{14}$$

Knowing that $m = 4.8748$ and $n = 2.4890$, we calculate the expectation $E(y)$ and the variance $V(y)$ from formulas (6) and (7). The results are $E(y) = 4.11807 + 0.24883x$, $V(y) = 1.23449$, and the standard deviation is $\sigma_y = 1.12191$. Fig. 3 plots Eq. (14) for the arbitrary value $x = 50$.

5. A CURRENT PARADIGM FREQUENTLY LEADS TO AN ERRONEOUS RISK PROFILE OF THE FORECASTED VARIABLE *Y* AND TO WRONGHEADED DECISIONS

First, recollect the possible loci the residuals can take. Earlier, we have explained the method for accurately calculating the risk profile (the accurate

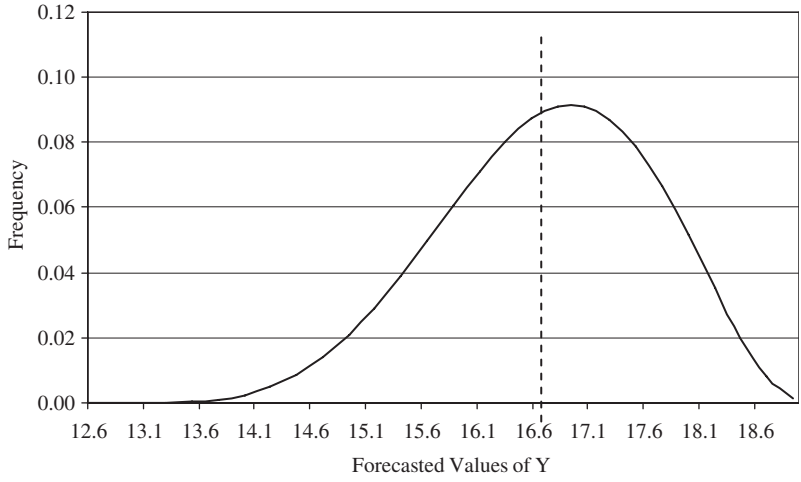

Fig. 3. The Density Function of Y. $E(y) = 16.63$ and $V(y) = 1.248$.

and complete PDF) for the random forecasted variable y. The risk profile will be either skewed to the left of its mode as in Fig. 4, or skewed to the right of its mode, as in Fig. 5, or it will be symmetric about its mean as in Fig. 6.

If y measures revenues, profits, return on investment, or other outcomes for which "large" is better than "small," then large values for y are more desirable than small values. In this case Fig. 5 offers the most attractive risk profile. If y measures costs, time, and resources expended, or other outcomes for which "small" is better than "large," then small values for y are more desirable than large values. In this case, Fig. 4 offers the most attractive risk profile. Fig. 6 is the risk profile of a "fair" wager that is biased to favor neither large nor small results.

Second, visit the current paradigm. Judging from the literature, which presents applications of regression forecasting models, a paradigm exists, which powerfully guides the approach that analysts follow. The paradigm rules are as follows:

- Measure the residual from a regression model;
- Measure the fit between the regression model's residuals pattern and a normally distributed pattern that obeys identical mean and standard deviation;

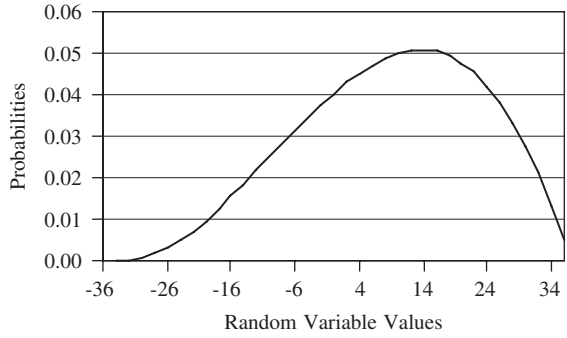

Fig. 4. Skewed to the Left of Mode.

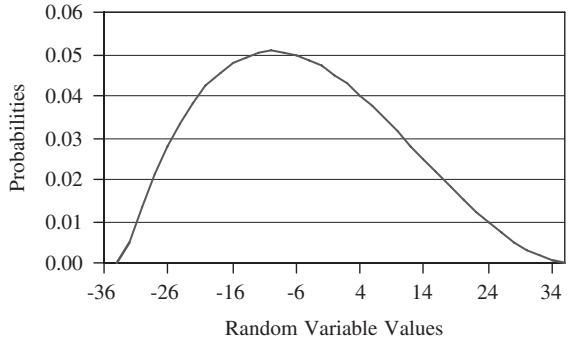

Fig. 5. Skewed to the Right of Mode.

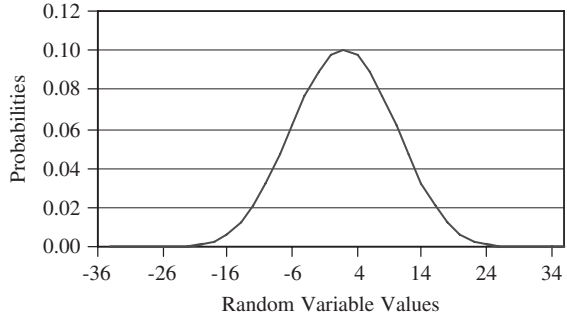

Fig. 6. Symmetric about the Average.

- Judge from the measured fit if the agreement between the two is "close enough." If the two are judged to be "close enough," the residual "can be considered" normally distributed; and
- The rest of the analysis of the regression model's forecasts will be based on statistical tests, confidence limits, and inferences from the normal distribution.

When the residual's histogram is symmetric about the mean, the paradigm may give wise guidance. Beware, however, that the mathematically astute knows that the normal density function does not fit very well many of the symmetric loci, which occur. But, when the residual's histogram is skewed, the paradigm guides the analysts to commit errors, often serious errors, and sometimes errors with disastrous consequences.

Third, we make our point through an application. In this section, we shall demonstrate the errors that can be committed when the regression's residual distribution is wrongly approximated by the normal function. We shall use as our standard the histogram of the residuals from regression model (1), graphed in Fig. 1 and correctly measured by the PPD function (10). The risk profile of y in model (1) is correctly measured by the PPD function (14). Therefore, Eqs. (10) and (14) are our reference points.

The normal approximation of the density of the forecasted variable y. The descriptive statistics of the residuals governed by the PPD function (10) are $E(\xi) = 0.0153$, $V(\xi) = 0.9239$, a standard deviation $\sigma_\xi = 1.12$, a lower bound $L_\xi = -4.00$, and an upper bound $M_\xi = 2.40$. The normal function fitted to these statistics is given by Eq. (15), except for the finite upper and lower bounds of ξ. The normal distribution has infinite bounds.

$$N(\xi|\mu_\xi, \sigma_\xi) = \frac{1}{1.12\sqrt{2\pi}} e^{(1/2)[(\xi-(-0.0135))/1.12]^2} \quad -\infty \le y \le +\infty \qquad (15)$$

Eq. (15) defines the normal approximation to the observed density of the residuals ξ. But it is the density of y we are interested in. To derive the density of the variable y from the known normal density (15) of the variable ξ, proceed as follows. Substitute the mean $E(y) = 4.1199 + 0.24883x$ of the regression model (1), and the model's standard deviation $\sigma_y = 1.12$ into (15). Multiply the result by the Jacobi transform $|J| = \partial\xi/\partial y = 1.00$, and obtain the normal density of y given by (16).

$$N(y|E(y), \sigma_y) = \frac{1}{1.12\sqrt{2\pi}} e^{(1/2)[(y-0.24883x-4.1199)/1.12]^2} \quad -\infty \le y \le +\infty \qquad (16)$$

For a given x, the mean of y is given by $E(y) = 4.11807 + 0.24883x$. For $x = 50$, we have $E(y) = 16.55$. Also, the variance $V(y) = 1.23449$, and the standard deviation is $\sigma_y = 1.12191$. Fig. 7 plots, in the same graph, both the correct PPD distribution of y (14) and the normal approximation (16) of y, for the arbitrary value $x = 50$ in both cases.

The ramifications from wrongly approximating the residuals revealed the following:

1. The intersections of the two curves in Fig. 7 reveal four intervals for y, which are of interest. The intervals are $12.0 \leq y \leq 14.7$, $14.7 \leq y \leq 17.0$, $17.0 \leq y \leq 18.6$, and $18.6 \leq y \leq 19.0$.
2. In first interval, $12.0 \leq y \leq 14.7$, the normal density and the PPD density calculate virtually identical probability estimates.
3. In the second interval, $14.7 \leq y \leq 17.0$, the normal approximation of the probabilities for the y values are higher than the true probabilities.
4. In the third interval, $17.0 \leq y \leq 18.6$, the normal approximation of the probabilities for the y values are smaller than the true probabilities.
5. In the fourth interval, $18.6 \leq y \leq 19.0$, the normal approximation of the probabilities for the y values are higher than the true probabilities.

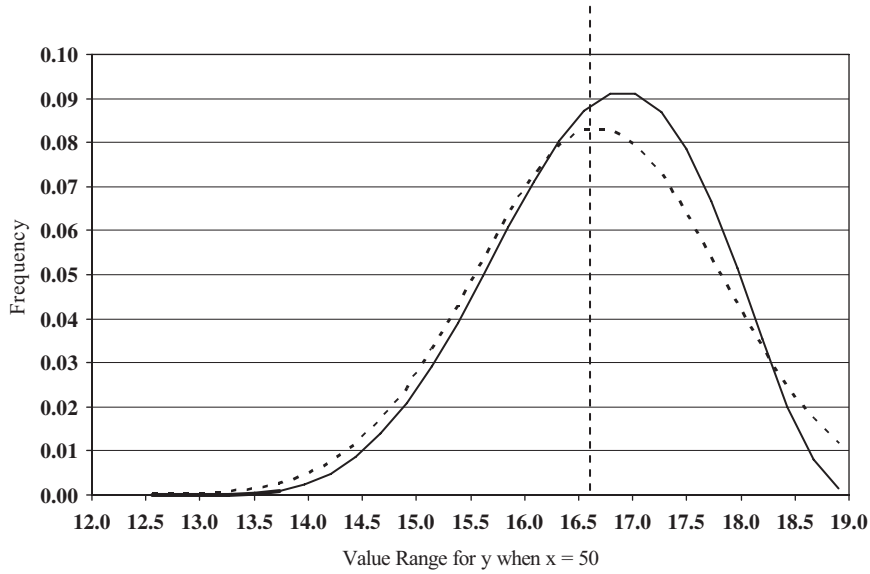

Fig. 7. The Probability Density of y by the Euler Density (bold) and by the Normal Density (dashed) when $x = 50$.

6. These discrepancies have ramifications. Suppose that a decision maker seeks wagers that offer high values of y with favorable probabilities (e.g., high Revenues, high profits, ROI). The normal approximation will be misleading the decision maker in intervals two, three, and four, and therefore for the wager as a whole.
7. Suppose that a decision maker seeks wagers that offer low values of y with favorable probabilities (e.g., low costs, low resource use). The normal approximation will similarly be misleading the decision maker in intervals two, three, and four.
8. The conclusions will be the same for a density function of y, which is skewed to the right.
9. Only in the case of symmetric densities for y will the normal approximation serve well, but not in all such cases.

Here then is a way to understand the ramifications that do not fit the residual density correctly can lead to. To give correct guidance to the decision makers, the policy makers, or the strategy makers, the correct density of y has to be known. It implies that the density of the residual's histogram must be correctly measured. This chapter shows that using the PPD family of functions will deliver highly accurate representations of the risk profile of the forecasted variable y.

6. CONCLUSION

Our objective was to develop a method for accurately calculating the risk profile of a forecasted variable. To accurately measure the risk profile of a forecasted variable y, it is necessary to first obtain an accurate representation of the histogram of a forecasting model's residual errors. The histogram of the residual ξ in a forecasting model $y = f(x) + \xi$ may be symmetric, or it may be skewed to either the left of or the right of its mode. To accurately represent the locus of any histogram, we introduced the PDF family of functions. That function is capable of correctly measuring the residual's histogram – be it skewed to the left, symmetric about the mean, or skewed to the right. From the residual's density, we correctly calculate the risk profile of the forecasted variable y from the density of the residual using the PPD function. We end up with the desired and accurate risk profile.

We discussed a universally followed paradigm that sometimes leads to misstating the risk profile and wrongheaded decisions. The paradigm guides

the analysts to approximate almost all histograms and densities with the Gauss–normal function. This chapter should open up many new avenues of progress for econometricians. The new PPD family of functions may stimulate many professional papers, each of which will have the power to develop econometric theory and applications further.

REFERENCES

Abramowitz, M., & Stegun, I. (Eds). (1972). *Handbook of mathematical functions with formulas, graphs and mathematical tables*. New York: Dover.

Allen, G., & Unwin, L. (1964). *The universal encyclopedia of mathematics*. New York: Simon and Schuster.

Arfken, G., & Weber, H. (2000). *Mathematical methods for physicists*. Harcourt: Academic Press.

Box, G. E. P., & Cox, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society B*, *26*, 211–243.

Brownlee, K. A. (1965). *Statistical theory and methodology in science and engineering* (2nd ed.). New York: Wiley.

Cleary, J., & Levenbach, H. (1982). *The professional forecaster: The forecasting process through data analysis*. Belmont, CA: Lifetime Learning Publications, A Division of Wadsworth, Inc.

Daniel, C., & Wood, F. S. (1999). *Fitting equations to data* (2nd ed). New York: Wiley-Interscience.

Davis, P. J. (1959). Leonard Euler's integral: A historical profile of the gamma function. *American Mathematical Monthly*, *66*, 849–869.

Feller, W. (1957). *An introduction to probability theory and its applications* (Vol. 1). New York: Wiley.

Feller, W. (1966). *An introduction to probability theory and its applications* (Vol. 2). New York: Wiley.

Granger, C. W. (1980). *Forecasting in business and economics*. New York, NY: Academic Press.

Graybill, F. A. (1961). *An introduction to linear statistical models*. New York, NY: McGraw-Hill Book Company.

Green, P. E., & Carroll, J. D. (1978). *Mathematical tools for applied multivariate analysis*. New York: Academic Press.

Haible, B., & Papanikolaou, T. (1997). *Fast multiprecision evaluation of series of rational numbers*. Technical Report No. TI-7/97. Darmstadt University of Technology, Germany.

Hair, J. F., Jr., Black, W. C., Babin, B. J., Anderson, R. E., & Tatham, R. L. (2006). *Multivariate data analysis*. Upper Saddle river, NJ: Pearson/Prentice-Hall.

Harris, R. J. (2001). *A primer of multivariate statistics* (3rd ed). Hillsdale, NJ: Lawrence Erlbaum Associates.

Havil, J. (2003). *Gamma, exploring Euler's constant*. ISBN 0-691-09983-9 (c).

Hildebrand, F. B. (1962). *Advanced calculus for applications*. Englewood Cliffs, NJ: Prentice-Hall.

Hines, W. W., & Montgomery, D. C. (1972). *Probability and statistics in engineering and management science* (3rd ed.). New York, NY: Wiley.
Kendall, M. G. (1951). *The advanced theory of statistics* (3rd ed., Vol. 2). New York, NY: Hafner Publishing Company.
Kendall, M. G., & Stuart, A. (1958). *The advanced theory of statistics* (Vol. 1).
Lebedev, N. N. (1972). *Special functions and their applications* (Translated and edited by R. A. Silverman). New York, NY: Dover Publications, Inc.
Levenbach, H., & Cleary, J. (1981). *The beginning forecaster: The forecasting process through data analysis*. Belmont, CA: Lifetime Learning Publications.
Maddala, G. S. (1977). *Econometrics*. New York: McGraw-Hill Book Company.
Makridakis, S., & Wheelwright, S. (1978). *Forecasting, methods and applications*. New York: Wiley.
Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression*. Reading, MA: Addison-Wesley.
Neter, J., Kutner, M. H., Wassermann, W., & Nachtsheim, C. J. (1996). *Applied linear regression models* (3rd ed.). Homewood, IL: Irwin.
Render, B., Stair, R. M., & Balakrishnan, R. (2006). *Managerial decision modeling with spreadsheets* (2nd ed.). Upper Saddle river, NJ: Prentice-Hall.
Render, B., Stair, R. M., & Hanna, M. (2006). *Quantitative analysis for management* (9th ed.). Upper Saddle river, NJ: Prentice-Hall.
Weisberg, S. (1985). *Applied linear regression*. New York: Wiley.
Wilks, S. S. (1962). *Mathematical statistics*. New York, NY: Wiley.
Wilson, E. B. (1912). *Advanced calculus*. (Chapter 14). Fort Dix, NJ: Ginn and Company.
Yamane, T. (1962). *Mathematics for economists, an elementary survey*. Englewood Cliffs, NJ: Prentice-Hall.

APPENDIX

Table A1. Empirical Observations on x and y in $y = ax + b$.

1	2	3	4	5	1	2	3	4	5
x	Observed y	Regression Estimates $y = 4.119 + 0.248x$	Residual ξ	Var(ξ)	x	Observed y	Regression Estimates $y = 4.119 + 0.248x$	Residual ξ	Var(ξ)
1	3.97	4.37	−0.40	0.16	38	14.42	13.57	0.85	0.72
2	5.82	4.62	1.20	1.45	39	13.87	13.82	0.05	0.00
3	5.47	4.87	0.61	0.37	40	14.32	14.07	0.25	0.06
4	2.72	5.11	−2.39	5.73	41	13.77	14.32	−0.55	0.30
5	6.37	5.36	1.01	1.02	42	15.42	14.57	0.85	0.72
6	5.42	5.61	−0.19	0.04	43	14.47	14.82	−0.35	0.12
7	6.67	5.86	0.81	0.66	44	14.17	15.07	−0.90	0.80
8	4.32	6.11	−1.79	3.20	45	15.77	15.32	0.45	0.21
9	6.17	6.36	−0.19	0.04	46	15.62	15.57	0.06	0.00
10	7.62	6.61	1.01	1.03	47	16.07	15.81	0.26	0.07
11	6.47	6.86	−0.39	0.15	48	15.92	16.06	−0.14	0.02
12	8.72	7.11	1.62	2.61	49	15.37	16.31	−0.94	0.89
13	6.57	7.35	−0.78	0.61	50	17.62	16.56	1.06	1.12
14	8.42	7.60	0.82	0.67	51	17.47	16.81	0.66	0.44
15	6.47	7.85	−1.38	1.91	52	16.92	17.06	−0.14	0.02
16	7.92	8.10	−0.18	0.03	53	16.37	17.31	−0.94	0.88
17	8.97	8.35	0.62	0.39	54	17.62	17.56	0.06	0.00
18	8.02	8.60	−0.58	0.33	55	18.27	17.80	0.47	0.22
19	8.87	8.85	0.02	0.00	56	19.32	18.05	1.27	1.61
20	10.32	9.10	1.23	1.50	57	17.57	18.30	−0.73	0.54
21	9.97	9.34	0.63	0.39	58	19.22	18.55	0.67	0.45
22	8.82	9.59	−0.77	0.60	59	18.27	18.80	−0.53	0.28
23	8.07	9.84	−1.77	3.14	60	20.72	19.05	1.67	2.80
24	9.52	10.09	−0.57	0.33	61	17.17	19.30	−2.13	4.52
25	11.17	10.34	0.83	0.69	62	20.02	19.55	0.47	0.22
26	10.02	10.59	−0.57	0.32	63	20.07	19.80	0.28	0.08
27	10.67	10.84	−0.17	0.03	64	21.12	20.04	1.08	1.16
28	10.32	11.09	−0.77	0.59	65	20.37	20.29	0.08	0.01
29	12.37	11.34	1.04	1.07	66	20.82	20.54	0.28	0.08
30	12.22	11.58	0.64	0.41	67	21.47	20.79	0.68	0.46
31	13.07	11.83	1.24	1.53	68	21.32	21.04	0.28	0.08
32	8.92	12.08	−3.16	9.99	69	22.17	21.29	0.88	0.78
33	11.92	12.33	−0.41	0.17	70	23.20	21.54	1.66	2.77
34	11.42	12.58	−1.16	1.34	71	22.27	21.79	0.48	0.23
35	14.27	12.83	1.44	2.08	72	18.12	22.04	−3.91	15.32
36	15.31	13.08	2.24	5.00	73	21.37	22.28	−0.91	0.83
37	11.77	13.33	−1.55	2.42	74	22.82	22.53	0.29	0.08

Table A1. (*Continued*)

1	2	3	4	5	1	2	3	4	5
x	Observed y	Regression Estimates $y = 4.119 + 0.248x$	Residual ξ	Var(ξ)	x	Observed y	Regression Estimates $y = 4.119 + 0.248x$	Residual ξ	Var(ξ)
75	23.27	22.78	0.49	0.24	90	24.62	26.51	− 1.89	3.58
76	22.72	23.03	− 0.31	0.10	91	28.27	26.76	1.51	2.27
77	21.97	23.28	− 1.31	1.71	92	28.12	27.01	1.11	1.23
78	22.62	23.53	− 0.91	0.82	93	25.77	27.26	− 1.49	2.22
79	23.33	23.78	− 0.44	0.20	94	28.42	27.51	0.91	0.83
80	25.32	24.03	1.30	1.68	95	28.27	27.76	0.51	0.26
81	24.37	24.27	0.10	0.01	96	27.72	28.01	− 0.29	0.08
82	25.22	24.52	0.70	0.49	97	28.37	28.26	0.11	0.01
83	24.67	24.77	− 0.10	0.01	98	29.62	28.50	1.12	1.25
84	26.32	25.02	1.30	1.69	99	25.47	28.75	− 3.28	10.78
85	26.17	25.27	0.90	0.81	100	29.32	29.00	0.32	0.10
86	26.02	25.52	0.50	0.25	101	28.17	29.25	− 1.08	1.17
87	25.27	25.77	− 0.50	0.25	102	31.02	29.50	1.52	2.31
88	26.32	26.02	0.30	0.09	103	28.47	29.75	− 1.28	1.63
89	25.17	26.27	− 1.09	1.20	104	30.52	30.00	0.52	0.27

Notes: 104 observations on x and y are presented in columns 1 and 2. Regressing x on y, using the least squares methods, produced the linear equation: $y = 4.119 + 0.2488x$. Predicted y values from this equation are shown in column 3. The residual ξ was calculated from $\xi = y - 0.2488x - 4.119$ and recorded in column 4. Also calculated from the data were $E(\xi) = -0.000135$, variance $\sigma^2 = V(\xi) = 1.23453$, and standard deviation is $\sigma = 1.12$.

This page intentionally left blank

FORECASTING USING INTERNAL MARKETS, DELPHI, AND OTHER APPROACHES: THE KNOWLEDGE DISTRIBUTION GRID

Daniel E. O'Leary

ABSTRACT

Much forecasting is done by experts, who either make the forecasts themselves or who do opinion research to gather such forecasts. This is consistent with previous knowledge management research that typically has focused on directly soliciting knowledge from those with greater recognized expertise.

However, recent research has found that in some cases, electronic markets, whose participants are not necessarily individual experts, often have been found to be more effective aggregated forecasters. This suggests that knowledge management take a similar tact and expand the perspective to include internal markets. As a result, this chapter extends the use of internal markets to be included in knowledge management, thus expanding the base of knowledge to gathering from nonexperts.

In particular, in this paper I examine the use of human expertise and opinion as a basis to forecast a range of different events. This chapter uses a "knowledge distribution grid" as a basis for understanding which kind of forecasting tool is appropriate for particular forecasting situations. We examine a number of potential sources of forecast information, including

Advances in Business and Management Forecasting, Volume 6, 157–172

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006010

knowledge acquisition, Delphi techniques, and internal markets. Each is seen as providing forecasting information for unique settings.

1. INTRODUCTION

Some forecasting questions, such as "Who will be elected President?" or "Who will win the Olympic Medal in Water Polo?" can use expert opinion, general opinion polls, or electronic markets as a basis of forecasting. Recent results have found that although expert opinion and opinion polls might receive the most publicity in the media, electronic markets are likely to provide a more accurate forecast of what will happen.

The same approaches might be used to address similar enterprise or corporate forecasting problems. As a result, it is probably not surprising that recent results have found that corporate internal markets provide insight that often is better than expert opinion.

However, having nonexperts, use internal virtual dollars to help develop forecasts, to assist corporations is a break from the classic approach, based on having experts forecast events.

1.1. Knowledge Management and Forecasting

Knowledge management systems generally gather knowledge from experts. For example, as seen in O'Leary (2008a, 2008b), the classic notion of expert systems and even new artifacts, such as Wikis, are based on the notion that some people know more than others and the knowledge management systems let them share that knowledge.

Further, historically, knowledge management has been backward looking, accumulating knowledge about what has occurred. For example, for consultants, knowledge management systems may gather proposals of previous engagements or summaries of actual engagements as key summaries of knowledge. Furthermore, other documents, such as news articles are likely to be accessible in such systems. Accordingly, virtually all the so-called knowledge management resources are backward looking. Knowledge is summarized for expert decision makers and they use that historical information to anticipate future events. The systems rarely provide forward-looking information, such as forecasts. Instead, experts use

the knowledge in the systems to generate forward-looking views and forecasts.

1.2. Internal Prediction Markets

However, recently enterprises and other organizations have begun to use internal markets to anticipate and forecast future events. For example (Wolfers & Zitzewitz, 2004), the Department of Defense was interested in knowing questions such as "Will the U.S. Military withdraw from country A in two years or less?"

Hewlett-Packard was one of the first companies to use such internal prediction markets (e.g., Totty, 2006). In 1996, they were concerned with how well such markets could forecast monthly printer sales. Approximately 15–20 people from various parts of the company were chosen to be a part of the market. Participants were given some cash and securities that were constructed to represent various monthly printer sales forecasts. In their market, only the winning sales number paid off to the participants. Using this internal market approach, the markets beat the experts 6 out of 8 times.

1.3. Expert vs. Nonexpert and Historic vs. Future

Accordingly, internal markets provide an approach that allows us to change the focus of knowledge management from just gathering knowledge from experts to a broader base of users (nonexpert). In addition, internal markets allow us to change our focus from a historical one to a view aimed at forecasting the future, rather than a historical view, summarizing the past. These results are summarized in Fig. 1.

1.4. Purpose of this Chapter

As a result, we need to understand those conditions under which to use alternative knowledge management approaches, particularly in forecasting of future events. Thus, this chapter is concerned with analyzing different approaches to gather human opinion and information as to the possible occurrence of future events.

Forecasting the answers to difficult problems often depends on asking the right person or group of people the right question. However, knowing which

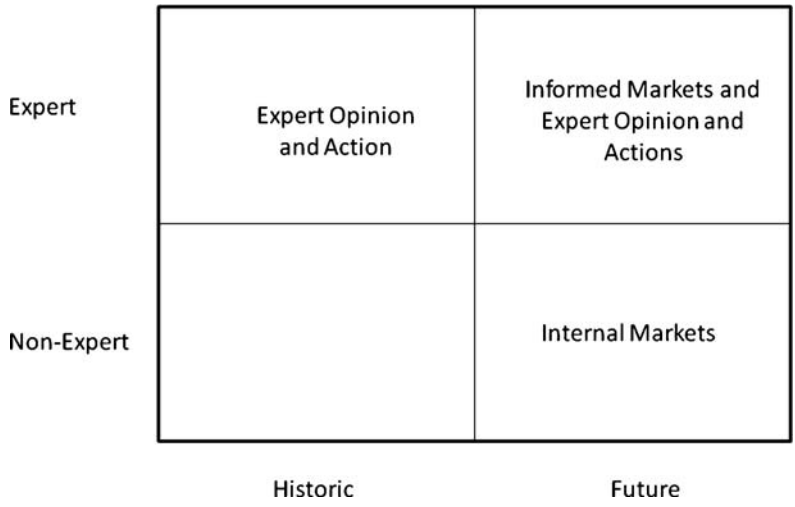

Fig. 1. Knowledge Management and Internal Markets.

approach to use is not always clear. Thus, the remaining purpose of this chapter is to outline two "knowledge distribution grids" that can be used to help determine which approach to forecasting is appropriate for particular situations, based on different characteristics of knowledge.

1.5. Outline of this Chapter

Section 2 examines notions of shallow knowledge vs. deep knowledge, whereas Section 3 examines distributed knowledge vs. concentrated knowledge. Section 4 examines a number of approaches used to gather knowledge in the different settings. Section 5 brings together Sections 2–4, and generates a knowledge distribution grid that allows us to better understand which approaches are useful for different forecasting opportunities in the cases of shallow and deep knowledge and distributed knowledge and concentrated knowledge. Section 6 extends the knowledge characteristics to dynamic and stable, and deterministic and probabilistic. Section 7 reviews some limitations of forecasting approaches, whereas Section 8 provides a brief summary of the chapter, and examines some contributions of the chapter and analyzes some potential extensions.

2. SHALLOW KNOWLEDGE VS. DEEP KNOWLEDGE

Different people may have knowledge about an area based on a range of factors, including ability, education, or experience. Further, knowledge may be distributed to a broad range or a small group of users. Thus, that knowledge may be distributed in varying depths to a range of users. The purpose of this section is to briefly discuss the first dichotomy, shallow and deep knowledge, and then examine some of the implications of that dichotomy.

2.1. Shallow Knowledge

In many cases, individuals have only shallow knowledge about particular issues. At the extreme, as noted by Hayek (1945, pp. 521–522)

> there is beyond question a body of very important but unorganized knowledge which cannot possibly be called scientific in the sense of knowledge of general rules: the knowledge of the particular circumstances of time and place ... (and that individuals have) ... special knowledge of circumstances of the fleeting moment, not known to others.

According to this description, there is asymmetric knowledge, particularly, in the case of knowledge that appears to describe events or contemporary activity. In particular, the knowledge is not general scientific knowledge. Further, that knowledge is distributed to a number of people, and the knowledge that people have may be relatively shallow.

In particular, knowledge is considered to be shallow if it is not connected to other ideas or if it is only loosely connected to other knowledge. Another view is that knowledge is shallow if it is more data than knowledge, or if that knowledge is or can be "compiled." For example, Chandrasekaran and Mittal (1999) suggested that if the knowledge can be put in a table (i.e., a classic table lookup) then that knowledge can be compiled and the corresponding knowledge is not particularly deep.

2.2. Deep Knowledge

Knowledge is regarded as "deep" if central or key issues in a discipline need to be understood to understand the issues at hand. Further, generally, knowledge has greater depth if it is connected to other ideas.

Deep knowledge also has been thought to be based on "first principles" (e.g., Reiter, 1987; Chandrasekaran & Mittal, 1999), rather than just based on causal knowledge. First principles provide a basis to reason from or about a set of issues. As an example, in the case of diagnostic systems, first principles ultimately employ a description of some system and observations of the behavior of the system and then reasons as to why the system may have failed.

Organizations are aware of and encourage cultivation of deep knowledge. For example, professional service firms have consulting "experts" in particular areas. A review of most consultants' resumes will rapidly tell you their areas of expertise. Universities are famous for having faculties that have deep knowledge in what can sometimes be very narrow areas. Oftentimes professional certifications can be issued to indicate that a level of depth and breadth in knowledge has been attained by an individual, for example, "certified public accountant" (CPA).

2.3. Shallow Knowledge vs. Deep Knowledge

One view of knowledge is that it comes in "chunks" (Anderson & Lebiere, 1998). Deep knowledge takes more "chunks" to be captured or described. Further, if there are numerous links between the chunks, rather than fewer, then that is another indication of deep knowledge. Thus, number of "chunks" provides one measure of the depth.

Research and other activities might change the classification by generating additional links or influencing the chunks. Potentially shallow knowledge may be made deeper if fragmented notions are connected or linked to other possibly related ideas. Alternatively, deep knowledge may be made shallower if it can be decomposed into relatively unconnected chunks.

2.4. Implications of Shallow vs. Deep Knowledge

One of the key tenets to system sciences (Ashby, 1956; Weick, 1969, p. 44) is that "it takes variety to destroy variety," or "it takes equivocality to remove equivocality." Thus, to paraphrase, "it takes shallow knowledge to capture shallow knowledge and deep knowledge to capture deep knowledge." If the knowledge of concern is shallow, rather than deep, then representation of the knowledge is likely to be easier. In addition, if the knowledge is shallow, then the basic approach used to capture shallow knowledge is likely to be

more straightforward. Further, capturing deep knowledge is likely to be more time consuming, and require greater resources.

2.5. Deep Knowledge and Internal Prediction Markets

In some settings deep knowledge is necessary to be able to attain an appropriate level of insight to a problem, even when using internal prediction markets. For example, if we are trying to forecast a flu epidemic this fall then knowledge of current viruses in circulation and the diffusion of those over time would be helpful at understanding the issue; otherwise, responses are likely to be little more than a random guess. Shallow knowledge of a problem, even if it is aggregated with knowledge of others is not likely to facilitate forecasting.

3. DISTRIBUTED KNOWLEDGE VS. CONCENTRATED KNOWLEDGE

In general, knowledge about forecasting events of interest can be what we call ''distributed'' among members of a population, such as an enterprise, or it can be concentrated in the hands of a relative few. The extent to which knowledge is distributed or concentrated can influence which approach is used to gather that knowledge.

3.1. Distributed Knowledge

Hayek (1945, p. 519) argues that much knowledge (in particular, knowledge of circumstances) is not ''concentrated,'' but instead is what we will call ''distributed,''

> The peculiar character of the problem of a rational economic order is determined precisely by the fact that the knowledge of the circumstances of which we must make use never exists in concentrated or integrated form, but solely as dispersed bits of incomplete and frequently contradictory knowledge which all of the separate individuals possess.

With some members knowing some things that others do not, this results in ''information asymmetry.'' Thus, with distributed knowledge, as noted by Hayek (1945, p. 520) ''knowledge (is) not given to anyone in its totality.''

3.2. Concentrated Knowledge

Hayek (1945) was specifically concerned with settings where a central group made decisions and set prices for the overall economy, rather than letting market forces determine prices. However, the concern generalizes to enterprises where, a central group of managers makes many decisions, rather than letting a broad range of users in the enterprise determine what should be done. Although some knowledge is concentrated in a few experts, much knowledge is distributed to a broad range of people in the enterprise. As a result, a critical characteristic of knowledge is whether knowledge is distributed in many different people or if knowledge is concentrated in a few specialists, and ultimately how the enterprise addresses this issue.

Concentrated knowledge refers to situations where expertise in a particular area is accomplished. There are numerous measures of concentration in different professional fields. For example, in accounting, there are numerous certifications, such as a CPA or certified management accountant (CMA). In education, various degree levels, for example, Ph.D., also denote concentration of knowledge.

In some economic systems, concentrated knowledge is assumed of central planners. In sporting events, odds makers (e.g., those from Las Vegas) are considered a concentrated source of knowledge. Other such assumed concentrated settings include planning departments or strategic planners.

3.3. Implications of Distributed or Concentrated Knowledge

In either case if we wish to manage the knowledge and make good decisions we need to determine the extent to which knowledge is either concentrated or distributed. The extent to which knowledge is concentrated or distributed influences the approaches that we will use to gather that knowledge. Thus, there are implications if the knowledge is distributed or concentrated.

If the knowledge is concentrated, then approaches to gather the knowledge would focus only on the small group with the resident knowledge. If we wish to be unobtrusive, then we could gather knowledge or implied knowledge from what people do. As a result, in the case of consultants, we could gather documents that consultants have generated and make those documents available to others, for example, engagement proposals and summaries. If the knowledge is concentrated into only a small group, then using an internal market to gather knowledge is not likely to be effective for various reasons discussed in the following text.

With distributed knowledge one must use approaches that tap into the asymmetric knowledge available to the many knowledgeable participants. If the knowledge is distributed, then gathering knowledge must take a different tact that tries to assemble the knowledge into an aggregated totality of sorts. In the case of forecasts, one such approach is internal markets, where we try to generate a solution that incorporates all of the disparate knowledge, through the choice of one alternative over another or an index representative of the choice concern.

3.4. Knowledge Diffusion

Knowledge is more likely to diffuse if there are so-called "network effects" that are dependent on communication and distributed knowledge. Thus, distributed knowledge seems more likely to diffuse, and more likely to diffuse more rapidly, than concentrated knowledge.

4. KNOWLEDGE GATHERING FOR FORECASTING

One of Hayek's (1945, p. 520) key concerns was with investigating "what is the best way of utilizing knowledge initially dispersed among all the people (as) is at least one of the main problems of economic policy – or of designing an efficient economic system." This chapter is basically concerned with the same issue. In particular, there are a number of ways to gather knowledge that could be used for forecasting, including the following.

4.1. "Man on the Street" Interview (One Person or More)

A well-known approach to gathering knowledge is to interview the random "man on the street," in an attempt to gather opinion or general knowledge of the populace. Such opinions could provide interviewers with an insight into a number of issues, such as would a particular product sell or who might win the presidential elections. Unfortunately, such interviews may or may not be successful at gathering the opinions desired, in part based on the limited sample, and the particular opinion.

The next step to generalization of "man on the street" interviews is to expand the sample size and gather opinion from a broader base of participants, and then aggregate their responses. Unfortunately, aggregation

is not an easy and noncontroversial issue. Given a base of opinions, how do we aggregate?

In general, opinion polls do not seek out those with any particularly deep expertise, but instead, seek a reasonable sample of some particular population. For example on college campuses, not surprisingly, frequently, the concern is with the opinion of the student population.

4.2. Knowledge Acquisition (1–10 People)

The development of the so-called "expert systems" and "knowledge-based systems" led to the analysis of what became known as knowledge acquisition, generally from a single expert (e.g., Rose, 1988). Enterprises generated a number of clever approaches to capture knowledge, such as videotaping events or the transfer of information between human actors (e.g., Kneale, 1986; Shpilberg, Graham, & Schatz, 1986) and those interested in being able to capture the knowledge.

However, over time there has been an interest in acquiring knowledge from multiple experts, rather than just a single expert. In the case of knowledge acquisition, "multiple" typically meant 2, 3, or 4, and rarely more than 10 (e.g., O'Leary, 1993; O'Leary, 1997). With multiple experts in such small samples, comes concerns as to whether the experts are from the same paradigm and whether the combination of expertise from different paradigms is sensible, and if it is sensible, how those multiple judgments should be combined.

4.3. Delphi Technique (5–30 People)

The Delphi technique (Dalkey, 1969; Green, Armstrong, & Graefe, 2007) is used to generate opinion, typically from expert. Using a three-step approach of gathering anonymous responses, controlling information feedback in a number of iterations, and aggregating opinions, the technique has been primarily used to generate consensus among a set of experts. The initial investigations used between 11 and 30 members in the group, but other investigations have used as few as 5. As a result, generally, this approach is used when the available expertise is distributed to a sufficiently large number of agents. In addition, this approach is used when there are time and resources to iteratively go back and forth with the experts to gradually elicit group consensus.

4.4. Enterprise or Internal Prediction Markets (20 or More People)

Enterprise markets, also known as "internal prediction markets" can be used to gather knowledge for forecasting a wide range of issues. Although internal prediction markets are not known as a knowledge management tool, in some cases they likely offer the best opportunity for gathering knowledge.

For example, recently internal prediction markets have been used to examine such issues as "Will our store in Shanghai open on time?" Such markets are "enterprise markets" since they are used by enterprise to forecast the future.

Internal prediction markets can be "informed markets" where the participants are those with more experience or knowledge about a particular area. Informed markets are necessary when the topic requires deep knowledge. For example, a market aimed at forecasting flu virus mutation likely necessarily would be an expert group.

5. A KNOWLEDGE DISTRIBUTION GRID

Our discussion about shallow knowledge vs. deep knowledge and distributed knowledge vs. concentrated knowledge, and knowledge gathering for forecasting is summarized in Fig. 2. Each axis has one of the two characteristics on it, yielding four different settings. It is referred to as a knowledge distribution grid because it provides view as to where knowledge is distributed and tools for gathering knowledge in different settings.

5.1. Gathering vs. Communicating

To this point, we have focused on gathering knowledge – whether dispersed or concentrated. However, in addition, the approaches provide differential communication devises between the sources of the knowledge and ultimately structuring the knowledge in a usable forecast form.

For example, when compared to markets, some authors think that Delphi may be easier to maintain confidentiality (e.g., Green et al., 2007) and that Delphi is more difficult to manipulate. Further, because Delphi is grounded in feedback, the approach is a bit more efficient than markets because other participants directly benefit from research of other participants that surface as feedback. This is in contrast with markets where participants each need to do their own research.

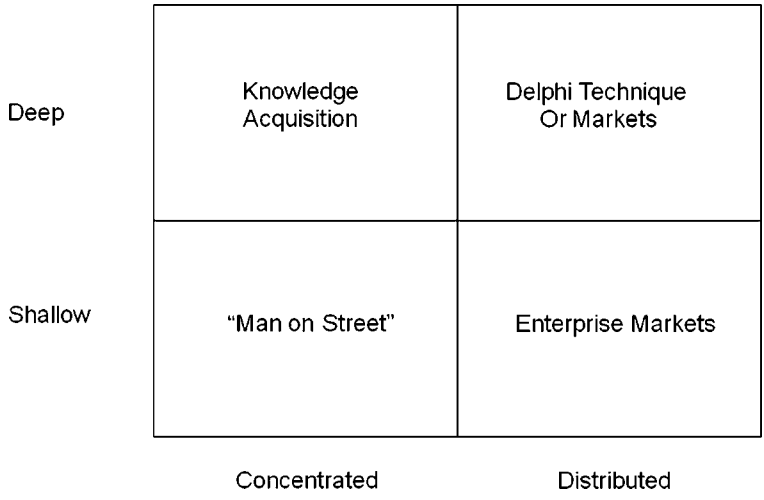

Fig. 2. Knowledge Distribution Grid.

6. EXTENSION TO DETERMINISTIC VS. PROBABILISTIC KNOWLEDGE AND STABLE VS. DYNAMIC KNOWLEDGE

Another set of characteristics that can be investigated include whether the knowledge is "stable vs. dynamic" or "deterministic vs. probabilistic." Knowledge may be stable over time or it may be dynamic, and change substantially as events unfold. However, knowledge may be deterministic or probabilistic. In some settings, conditions and events appear deterministically related. For example, many events in electrical or plumbing systems are often characterized as deterministic. In those settings, given a set of conditions, deterministic forecasts about what will happen can be made. Alternatively, in other systems events are more probabilistic. These dimensions are summarized in Fig. 3.

If we suppose that dynamic is more complex than stable and probabilistic is more complex than deterministic, then dynamic and probabilistic systems are the most complex of all four conditions. As a result, in general, as we move out of the lower left quadrant, and into the upper right quadrant, events become more complex.

How knowledge fits in these categories can impact which approach can be used to forecasts events. In quadrant #1, where there are relatively stable and

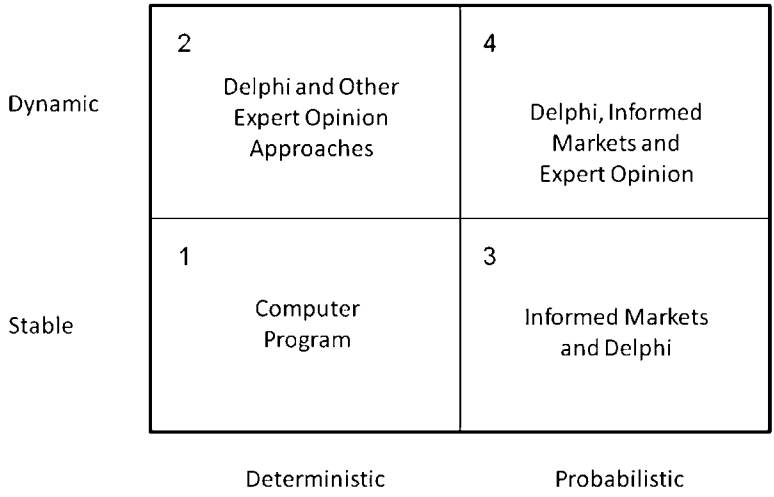

Fig. 3. Dynamic vs. Stable and Deterministic vs. Probabilistic.

deterministic problems there often are sets of rules that forecast an outcome, for example, credit determination. In this setting, a computer program can be used that embodies those rules to provide a consistent and cost-effective approach to forecast when knowledge characteristics fit here.

In quadrant #2, where the world is dynamic and fast changing, it takes a unique set of experts to keep up with the events to forecast the future. As a result, expert opinion is likely to be particularly helpful in this setting.

In quadrant #3, the knowledge is stable, but probabilistic. In this setting, a market can be executed and a probabilistic estimate gathered. Knowledge is stable enough to allow a market to evolve.

Quadrant #4 is the most difficult because knowledge is dynamic and probabilistic. If the knowledge changes too rapidly, then before a market of informed participants can be successfully generated, the solution may have changed. However, if it is too dynamic, then perhaps expert opinion is the answer. As a result, Delphi, done in a time manner, could provide appropriate insights. In this quadrant, there is no one solution approach.

7. CONCERNS

There are some potential sample concerns, no matter which approach is used to gather potential forecasts. First, the size of the sample from which

the knowledge and forecasts is gathered is critical to generating an appropriate view of the future. With internal markets, this issue is referred to under the notion of "thin markets" where there are too few traders to guarantee an effective and efficient market. Probably not unexpectedly, when using internal markets, in general, larger markets are better. However, if the sample size gets too large, then Delphi can bog down.

Second, sample bias from the population investigated has been considered by researchers in most of these disciplines. In general, the smaller the population is being drawn from, the bigger is the potential problem of bias. However, in some of the approaches discussed earlier, for example, in the case of knowledge acquisition, the problem has been largely ignored.

Third, there is the potential for conflicts between points of view. In the arena of knowledge acquisition, O'Leary (1993, 1997) has investigated this issue. As part of the concern of knowledge acquisition from multiple experts, there has been concern with the so-called "paradigm myopia." In that setting, with few experts as the basis of the knowledge used, any bias is likely to be reflected in the forecast that is made. Similarly, when using Delphi, a diverse opinion set can drive the group to generate unique solutions or deadlock them.

8. SUMMARY, CONTRIBUTIONS, AND EXTENSIONS

This chapter has differentiated characteristics of knowledge along a number of dimensions and investigated those dichotomous characteristics, including

- expert vs. non-expert and historic vs. future,
- deep vs. shallow and concentrated vs. distributed,
- dynamic vs. stable and deterministic vs. probabilistic.

Using those characteristic pairs, we analyzed settings where different approaches were more appropriate than others. In particular, we generated the knowledge distribution grid, and used the grid to differentiate between different approaches that might be used for forecasting future events. Four different basic approaches were discussed: "man-on-the-street," "knowledge acquisition," Delphi technique, and internal prediction markets. Markets were labeled as "informed" for those involving experts, and "enterprise" for those internally conducted by enterprises to address issues of direct concern to the enterprise.

8.1. Contributions

This chapter has extended knowledge management to be forward looking and to include internal markets as a means of gathering knowledge from a broad base of users. Virtually all previous knowledge management has focused on knowledge management as a medium to capture historical information. Further, internal markets have been ignored as a knowledge management tool. In addition, comparison of internal markets to approaches such as Delphi, have received little attention.

This chapter provides three distinct sets of grids featuring different knowledge characteristics. Each of those grids provides the ability to investigate knowledge characteristics, and how those knowledge characteristics influence different approaches to gathering information for forecasting future events.

8.2. Extensions

This research can be extended in a number of directions. First, we could expand the number of methods examined in Section 4. For example, other approaches, such as using surveys to gather knowledge and opinions could be generated. Second, the discussion in this chapter has suggested that the time and resources available also influence the choice of an approach to forecast the future. Compiled expertise would be a rapid approach, but running an internal market while investigating time constrained issues is not likely to provide timely returns. Third, in some cases there appear to be multiple feasible approaches to problems. For example, Delphi and internal markets both seem feasible with approximately 20 or more people and an expert environment. However, it is not clear which approaches provide the "best results." As a result, the two approaches could be compared for which works best under which conditions, and the strengths and weaknesses of each could be more fully fleshed out.

REFERENCES

Anderson, J., & Lebiere, C. (1998). *The atomic components of thought*. Mahwah, NJ: Lawrence Erlbaum and Associates.

Ashby, W. R. (1956). *An introduction to cybernetics*. New York: Wiley.

Chandrasekaran, B., & Mittal, S. (1999). Deep versus compiled knowledge approaches to diagnostic problem solving. *International Journal of Human Computer Studies*, *51*, 357–368.

Dalkey, N. C. (1969). *The delphi method: An experimental study of group opinion, June*. Rand Report RM-5888-PR. Available at http://www.rand.org/pubs/research_memoranda/2005/RM5888.pdf

Green, K. C., Armstrong, J. S., & Graefe, A. (2007). *Methods to elicit forecasts from groups: Delphi and prediction markets compared, August 31*. MPRA Paper 4663.

Hayek, F. (1945). The use of knowledge in society. *The American Economic Review*, *XXXV*(September 4), 519–530.

Kneale, D. (1986). How Coopers & Lybrand put expertise into its computers. *The Wall Street Journal* (November 14).

O'Leary, D. E. (1993). Determining differences in expert judgment. *Decision Sciences*, *24*(2), 395–407.

O'Leary, D. E. (1997). Validation of computational models based on multiple heterogeneous knowledge sources. *Computational and Mathematical Organization Theory*, *3*(2), 75–90.

O'Leary, D. E. (2008a). Expert systems. In: B. Wah (Ed.), *Encyclopedia of computer science*. New York, USA: Wiley.

O'Leary, D. E. (2008b). Wikis: From each according to his knowledge. *Computer*, *41*(2), 34–41.

Reiter, R. (1987). A theory of diagnosis from first principles. *Artificial Intelligence*, *32*, 57–95.

Rose, F. (1988). An 'electronic' clone of a skilled engineer is very hard to create. *The Wall Street Journal* (August 12).

Shpilberg, D., Graham, L., & Schatz, H. (1986). Expertax: An expert system for corporate tax accrual and planning. *Expert Systems*, *3*(3), 136–151.

Totty, M. (2006). Business solutions. *The Wall Street Journal* (June 19), R9.

Weick, K. (1969). *The social psychology of organizing*. Addison-Wesley: Reading, MA.

Wolfers, J., & Zitzewitz, E. (2004). Prediction markets. *Journal of Economic Perspectives*, *18*(2), 107–126.

THE EFFECT OF CORRELATION BETWEEN DEMANDS ON HIERARCHICAL FORECASTING

Huijing Chen and John E. Boylan

ABSTRACT

The forecasting needs for inventory control purposes are hierarchical. For stock keeping units (SKUs) in a product family or a SKU stored across different depot locations, forecasts can be made from the individual series' history or derived top–down. Many discussions have been found in the literature, but it is not clear under what conditions one approach is better than the other. Correlation between demands has been identified as a very important factor to affect the performance of the two approaches, but there has been much confusion on whether it is positive or negative correlation. This chapter summarises the conflicting discussions in the literature, argues that it is negative correlation that benefits the top–down or grouping approach, and quantifies the effect of correlation through simulation experiments.

INTRODUCTION

Many organisations operate in a multi-item, multi-level environment. In general, they have to "cope with well over 100 time series with numbers over

Advances in Business and Management Forecasting, Volume 6, 173–188

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006011

10,000 being quite common'' (Fildes & Beard, 1992). These time series are often related. For example, a company may group similar products in product families according to specifications, colours, sizes, etc. Alternatively, in a multi-echelon inventory system, a stock-keeping unit's sales may be recorded in many different locations at varying levels of aggregation. Therefore, in such cases, the data available and the need for forecasts are hierarchical.

A substantial part of the forecasting literature has been devoted to models and methods for single time series. However, as indicated earlier, the short-term forecasting need for production and inventory control purposes is to address a large amount of series simultaneously. Duncan, Gorr, and Szczypula (1993) argued that ''forecasting for a particular observational unit should be more accurate if effective use is made of information, not only from a time series on that observational unit, but also from time series on similar observational units.''

There have been many discussions on group forecasting in the literature. However, no clear conclusions have been reached on the conditions under which the grouping approach is better than the individual approach. Correlation between demands has been identified as a very important factor, but there has been much confusion about whether positive or negative correlation would benefit grouping. This chapter is presented as follows: contrasting arguments are discussed in the next section; then the findings on the role of correlation from simulation experiments are presented; and, finally, the chapter is summarised with our conclusions. The overall purpose of this chapter is to dispel some of the confusion in the literature on how correlation affects the grouping approach.

DEBATES AND CONFUSION IN THE LITERATURE

It is well recognised that to obtain better forecasts, one should make better use of available forecasting series. Some practitioners such as Muir (1983), McLeavey and Narasimhan (1985), and Fogarty and Hoffmann (1983) have argued that forecasting an aggregate and then allocating it to items is more accurate than generating individual forecasts. Their argument was that the top–down approach resulted in more accurate predictions since aggregate data were more stable.

Schwarzkopf, Tersine, and Morris (1988) pointed out two problems of using the top–down approach: model incompleteness and positive correlation. They argued that the aggregate model may not completely describe the

processes in the individual series, that is, there were model differences among the series. When the total forecast was disaggregated back to the item level, correlated errors were produced. They commented that "this modelling error can be quite large and may override the more precise potential of top–down forecasts" (Schwarzkopf et al., 1988). The same point was also made by Shlifer and Wolff (1979). The second problem was that if there was a strong positive correlation in demand for items in a group, the variance for the aggregate was increased by the amount of the covariance term. Schwarzkopf et al. (1988) advanced our understanding of some of the reasons why the top–down approach does not always lead to a more accurate subaggregate forecast.

Top–down forecasts have to be treated with caution. If the individual series follow different demand generating processes, then the aggregate model does not reflect any of those individual processes. Although the aggregate data are less noisy, it does not always result in more accurate subaggregate forecasts. Even when the modelling difference is not an issue, there is an additional problem of the disaggregation mechanism to be applied.

One way to get around these problems is to group seasonal homogeneous series. From a classical decomposition point of view, demand consists of level, trend, seasonality and noise. In a group of items, levels can be varying. Trends can be upwards or downwards and can have various degrees. However, seasonality is often more stable as it is affected by weather and customs. Chatfield (2004) pointed out that seasonal indices are usually assumed to change slowly through time, so that $S_t \approx S_{t-q}$, where q is the seasonal cycle. It makes more sense to use the grouping approach to estimate seasonality than to estimate level and trend as there is an issue of modelling difference. The problem of an appropriate disaggregation mechanism can also be avoided. For multiplicative seasonality, no disaggregation mechanism is needed as seasonality is relative to the mean. For an additive model with common seasonal components across the group, a simple average can be used as the disaggregation method. Although it is difficult to apply the grouping approach in general, we found it helpful in seasonal demand forecasting, that is, estimating level and trend individually but seasonality from the group.

Correlation has been identified as a very important factor to affect the grouping and individual approaches, but there has been some confusion about whether positive or negative correlation benefits grouping. Duncan, Gorr, and Szczypula (1998) argued for positive correlation. They claimed that analogous series should correlate positively (co-vary) over time.

Then the co-variation would be able to "*add precision to model estimates and to adapt quickly to time-series pattern changes*". However, Schwarzkopf et al. (1988) supported negative correlation as the covariance term was increased by positive correlation. The confusion lies in the distinction between a common model and varied models. Given the same model, it is negative correlation between series that reduces variability of the total and favours the top–down approach. However, the more consistent the model forms are, the more this favours the grouping approach; and consistency of model forms is associated with positive correlations between series, not negative correlations. Duncan et al. (1998) also identified the association between consistency of model forms and positive correlations. However, positive correlations should not be used to identify whether different series follow the same model, as sometimes the positive correlations may be incurred by a trend component, rather than the model form. Therefore, checks should be made on the consistency of models using other diagnostics, before employing correlation analysis to establish whether a grouped or individual approach is preferable.

SIMULATION EXPERIMENTS

We used simulation experiments to examine the effect of correlation between demands on forecasting performance of the grouping and individual approaches. It is argued that it is negative correlation that will benefit the grouping approach when a common model is assumed; the main purpose of the simulation experiments is to quantify the effect.

Two simple models are assumed to generate demand

$$Y_{i,th} = \mu_i S_h + \varepsilon_{i,th} \tag{1}$$

$$Y_{i,th} = \mu_i + S_h + \varepsilon_{i,th} \tag{2}$$

where i is a suffix representing the stock keeping unit (SKU) or the location, suffix t the year and $t = 1, 2, \ldots, r$ (where r is the number of years' data history), suffix h the seasonal period and $h = 1, 2, \ldots, q$ (where q is the length of the seasonal cycle), Y the demand, μ_i the underlying mean for the ith SKU or location and is assumed to be constant over time but different for different SKUs or locations, S_h a seasonal index at seasonal period h (it is unchanging from year to year and the same for all SKUs or locations under consideration), $\varepsilon_{i,th}$ a random disturbance term for the ith

SKU/location at the tth year and hth period (it is assumed to be normally distributed with mean zero and constant variance σ_i^2). There are correlations ρ_{ij} between $\varepsilon_{i,th}$ and $\varepsilon_{j,th}$ at the same time period. Auto-correlations and correlations at different time periods are assumed to be zero.

Model (1) has multiplicative seasonality and model (2) has additive seasonality. It is assumed that there is no trend so that we may focus on the seasonality. The underlying mean is assumed to be stationary. Seasonality is also assumed to be stationary and the same within the group.

Trend components are not considered in the current models to avoid the complexity of different degrees of trend when aggregating, and thus focus mainly to gain insights into the effect of correlation from simple models.

The estimator for the underlying mean is

$$\hat{\mu}_i = \frac{1}{qr}\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{i,th} \tag{3}$$

The individual seasonal indices (ISI) estimator for the mixed model (Eq. (1)) is

$$\hat{S}_h = \mathrm{ISI}_{i,h} = \frac{q\sum_{t=1}^{r} Y_{i,th}}{\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{i,th}} \tag{4}$$

The ISI estimator for the additive model (Eq. (2)) is

$$\hat{S}_h = \mathrm{ISI}_{i,h} = \frac{1}{r}\sum_{t=1}^{r} Y_{i,th} - \frac{1}{qr}\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{i,th} \tag{5}$$

Two group seasonal indices (GSI) methods have been proposed from the literature. Dalhart (1974) proposed a method that was a simple average of the ISI.

$$\mathrm{DGSI}_h = \frac{1}{m}\sum_{i=1}^{m} \mathrm{ISI}_{i,h} = \frac{q}{m}\sum_{i=1}^{m} \frac{\sum_{t=1}^{r} Y_{i,th}}{\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{i,th}} \tag{6}$$

Withycombe (1989) suggested aggregating all the individual series first and then estimating seasonal indices from the aggregate series

$$\mathrm{WGSI}_h = \frac{q\sum_{t=1}^{r} Y_{A,th}}{\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{A,th}} \tag{7}$$

Both DGSI and WGSI were proposed to multiplicative seasonality. When seasonality is additive, the two methods are the same and we call it GSI.

$$\mathrm{GSI}_h \frac{1}{mr}\sum_{t=1}^{r} Y_{A,th} - \frac{1}{mqr}\sum_{t=1}^{r}\sum_{h=1}^{q} Y_{A,th} \tag{8}$$

We have developed rules to choose the best method between the ISI and GSI methods. Interested readers can refer to Chen and Boylan (2007).

SIMULATION FINDINGS

The simulation results quantify the effect of correlation on the forecasting performance of the individual and grouping approaches. We use mean square error (MSE) as the error measure and report the percentage best (PB) results.

Results for the additive model are presented first, followed by results for the mixed model. Detailed simulation designs are presented in Appendix A.

Table 1 shows that negative correlation favours GSI. As the correlation coefficient changes from highly negative to highly positive, the number of series for which GSI is the best decreases. This is consistent with the theory that as correlation changes from highly negative to highly positive, ISI will

Table 1. Effect of Correlation on the Percentage of Series for Which ISI or GSI is the Best (Additive Model).

Correlation	−0.9	−0.6	−0.3	0	0.3	0.6	0.9
ISI (%)	40.00	41.43	43.84	45.63	48.57	50.00	50.00
GSI (%)	60.00	58.57	56.16	54.38	51.43	50.00	50.00

Table 2. Effect of Correlation on the Percentage of Series for Which ISI or GSI is the Best (Mixed Model).

Correlation	−0.9	−0.6	−0.3	0	0.3	0.6	0.9
ISI (%)	0.00	0.00	0.00	1.13	13.96	29.55	48.78
DGSI (%)	64.11	52.92	46.67	40.65	27.35	11.16	5.71
WGSI (%)	35.89	47.08	53.33	58.21	58.69	59.29	45.51

be the best for more series. When the correlation coefficient is 0.6 or 0.9, ISI and GSI are equally good.

Table 2 shows that for the mixed model, ISI is never the best when correlation is negative. DGSI is the best when correlation is highly negative (between −0.9 and –0.6), and the number of series for which DGSI is the best decreases when correlation increases. The number of series for which WGSI is the best increases as correlation increases. But for a very high positive correlation coefficient of 0.9, ISI becomes the best method. Simulation results clearly show that GSI is better than ISI for the majority of possible correlations within the range.

The case of two series is simplistic, although it provides useful insight into the effect of correlation. In reality, the number of items in a group can be as large as hundreds or even thousands. To cover more realistic situations, we now proceed to simulate groups of more than two series (detailed design can be found in the Appendix A).

The group size has to be determined somewhat arbitrarily. In this simulation, we define the group size to be $2^n (n = 1, 2, 3, \ldots, 6)$. So the group sizes are 2, 4, 8, 16, 32 and 64. The group size increases with an unequal and growing increment because when the group size is smaller, we want to examine the effect at a finer level. When the size is larger, it is increasingly difficult to do so. The maximum group size is 64 because of the time and complexity of computing.

Correlation coefficients cannot be decided arbitrarily as in the case of a group of two series, nor can the correlation matrix be generated randomly. A feasible correlation matrix must be positive semi-definite, that is, all the eigenvalues must be non-negative (see, e.g., Xu & Evers, 2003). Therefore, we followed the algorithm suggested by Lin and Bendel (1985) to generate feasible matrices with specified eigenvalues.

Ideally we would like to cover a comprehensive set of correlation matrices, but the number of possible combinations of feasible matrices makes this impossible. Instead, one looks at a number of feasible correlation matrices covering a range as large as possible.

Correlation does not affect the ISI method. For DGSI, it is $\Sigma_{j=1}^{m-1}\Sigma_{l=j+1}^{m}((1/\mu_j)(1/\mu_l))\rho_{jl}\sigma_j\sigma_l$ that matters. For WGSI, it is σ_A^2, which equals $\sigma_1^2+\sigma_2^2+\cdots+\sigma_m^2+2\Sigma_{j=1}^{m-1}\Sigma_{l=j+1}^{m}\rho_{jl}\sigma_j\sigma_l$. It is the term $\Sigma_{j=1}^{m-1}\Sigma_{l=j+1}^{m}\rho_{jl}\sigma_j\sigma_l$ that involves the correlation coefficients (Chen & Boylan, 2007). However, for both DGSI and WGSI, it is not straightforward from the theoretical expressions how correlation affects the rules. The standard deviations (coefficients of variation in DGSI) are interacting with the correlation coefficients and cannot be separated. What we want to see is what structure of correlation matrix affects the rules, and we will do this by calculating the lower and upper bounds of the cross terms (details can be found in Appendix B).

Let $P^+=\Sigma_{i=1}^{m-1}\Sigma_{j=i+1}^{m}\rho_{ij}^+$ and $P^-=\Sigma_{i=1}^{m-1}\Sigma_{j=i+1}^{m}\rho_{ij}^-$; for simulation purposes, we can experiment with different values of P^+ and P^- to evaluate the effect of correlation. With the bounds, the σ_i terms are separated from the correlation coefficients. However, in reality the cross term is not a simple function of the correlation coefficients but the interaction of correlation coefficients and the standard deviation terms. For given σ_i terms, the cancellation depends not only on the values of P^+ and P^- but also on the positions of the positive and negative coefficients. From a simulation perspective, it is difficult to experiment with both sign and position of each correlation coefficient. Therefore, we bring the problem down to the two dimensions of P^+ and P^-.

We will generate 1,000 different feasible correlation matrices for each group size n. It is a very small proportion of all possible combinations of feasible correlation matrices. We cannot use all of these feasible correlation matrices in our simulation to examine the effect of correlation along with other parameters. Just as we vary all the other parameters that affect the rules, we will vary P^+ and P^- too. Out of the 1,000 feasible correlation matrices we generate, we will calculate $|P^+/P^-|$ and then choose the minimum, the first quartile, the second quartile (median), the third quartile and the maximum. This covers the whole range of the correlation matrices we generated. Then these five matrices are used in the simulations and their interactions with other parameters can be assessed. Table 3 shows the range of $|P^+/P^-|$ for each group size.

When the additive model is assumed, GSI outperformed ISI universally. Therefore, we cannot analyse the effect of the different correlation matrices on ISI and GSI. However, the effect is analysed for the mixed model in Table 4.

For each group size, five different correlation matrices are chosen in our simulation experiments according to different ratios of $|P^+/P^-|$. Matrix 1

Table 3. Range Ratio of Positive and Negative Correlation Coefficients.

Group Size	4	8	16	32	64
Minimum	0.0000	0.0847	0.2313	0.4436	0.6497
Lower quartile	0.1674	0.4142	0.5719	0.6851	0.7592
Median	0.3500	0.5715	0.6919	0.7637	0.8260
Upper quartile	0.6267	0.7639	0.8287	0.8678	0.9031
Maximum	2.7847	1.4352	1.0817	1.0217	0.9975

Table 4. Effect of Correlation Matrix on the Percentage of Series for Which DGSI or WGSI is the Best.

Correlation Matrix	1	2	3	4	5
DGSI (%)	76.05	52.90	50.99	45.48	41.97
WGSI (%)	23.95	47.10	49.01	54.52	58.03

has the lowest $|P^+/P^-|$ and matrix 5 has the highest $|P^+/P^-|$. ISI is never the best. When $|P^+/P^-|$ increases, the percentage of series for which DGSI is the best decreases and the percentage of series for which WGSI is the best increases. This is what we expected. Simulation results from group of two series show that DGSI was the best when correlation was between –0.9 and –0.6, WGSI was the best when correlation was between –0.3 and 0.6 and beyond that ISI became the best. Therefore, the greater the sum of all negative correlation coefficients, the more series for which DGSI would be expected to be the best.

Previous research on the issue of grouping has consistently suggested correlation as the most important factor to decide whether a direct forecast or a derived (top–down) forecast should be used. However, there have been arguments on whether series with positive or negative correlation favours the derived approach. Our simulation results reveal that for a wide range of positive correlation values, GSI methods are still better than the ISI method; but the gain of using the GSI methods is greater when series are negatively correlated.

Our simulation of two series in a group is much more specific than previous research: it does not only show the range of correlation that a GSI method outperforms the ISI method, but it also shows the range of correlation for which one GSI method outperforms the other. Within the former range (-0.9 to 0.6 in our simulation), DGSI outperforms WGSI

when correlation is between –0.9 and –0.6 and WGSI is better when correlation is between –0.3 and 0.6. It is not until correlation is almost as high as 0.9 that ISI becomes the best performing method.

When there are more than two series in the group, it is more difficult to find clear cut how correlation affects the individual and grouping approaches. Our simulations of up to 64 items in a group and five different correlation matrices show that ISI is never better than the grouping approach. Moreover, we found that DGSI is better for lower $|P^+/P^-|$ and WGSI is better for higher $|P^+/P^-|$. This is consistent with the findings in the case of two series.

EXTENSION TO MODELS WITH TREND

The current models we assume are simple ones without a trend component. A key finding is that correlation between demands is induced only by correlation between the error terms in the model.

Take the additive model $Y_{i,th} = \mu_i + S_h + \varepsilon_{i,th}$
The deseasonalised demand is $Y^*_{i,th} = Y_{i,th} - \hat{S}_h$
Since $\hat{S}_h$ (Eq. (5)) is an unbiased estimator, $\mathrm{E}(\hat{S}_h) = S_h$

$$\begin{aligned}\mathrm{cov}(Y^*_{i,th}Y^*_{j,th}) &= E\left\{[Y_{i,th} - \hat{S}_h - E(Y^*_{i,th})][Y_{j,th} - \hat{S}_h - E(Y^*_{j,th})]\right\} \\ &= E\left[(\mu_i + S_h + \varepsilon_{i,th} - \mu_i - S_h)(\mu_j + S_h + \varepsilon_{j,th} - \mu_j - S_h)\right] \\ &= E(\varepsilon_{i,th}\varepsilon_{j,th}) \qquad (9)\end{aligned}$$

Therefore, the only source of correlation between demands is from correlation between the random error terms.

We can extend the analysis to an additive trend and seasonal model.

Assume

$$Y_{i,th} = \mu_i + [(t-1)q + h]\beta_i + S_h + \varepsilon_{i,th} \qquad (10)$$

where $[(t-1)q + h]\beta_i$ is the trend term and β_i the growth rate.

Suppose we can find an estimator $\hat{\beta}_i$ for β_i, then to detrend the model we have

$$Y_{i,th} - [(t-1)q + h]\hat{\beta}_i = \mu_i + [(t-1)q + h](\beta_i - \hat{\beta}_i) + S_h + \varepsilon_{i,th} \qquad (11)$$

The detrended and deseasonalised demand is

$$Y^*_{i,th} = Y_{i,th} - [(t-1)q + h]\hat{\beta}_i - S_h \tag{12}$$

Therefore, assuming $\beta_i - \hat{\beta}_i$ is independent of $\beta_j - \hat{\beta}_j$ and $\beta_i - \hat{\beta}_i$ is independent of $\varepsilon_{j,th}$

$$\begin{aligned}\operatorname{cov}(Y^*_{i,th}Y^*_{j,th}) &= E\left[(Y^*_{i,th} - \mu_i)(Y^*_{j,th} - \mu_j)\right] \\ &= E\left\{([(t-1)q+h](\beta_i - \hat{\beta}_i) + \varepsilon_{i,th})([(t-1)q+h](\beta_j - \hat{\beta}_j) + \varepsilon_{j,th})\right\} \\ &= E(\varepsilon_{i,th}\varepsilon_{j,th}) \end{aligned} \tag{13}$$

This result shows that correlation between the demands is induced only by correlation between the error terms in the model. This is the same as Eq. (9); so the same result carries through from a non-trended model to a trended model.

This same approach does not apply for the mixed model though. It requires a different approach to investigate the effect of correlation assuming a multiplicative trend and seasonality model.

Further research can also extend beyond stationary seasonality and consider time-varying Winters' type models. This line of research is undertaken by another group of researchers (Dekker, van Donselaar, & Ouwehand, 2004; Ouwehand, 2006). They derived a model to underlie the multivariate version of the Holt–Winters' method, that is, estimating level and trend individually and seasonal indices from the group. However, the effect of correlation is yet to be addressed.

CONCLUSIONS

This chapter clarifies some of the confusion in the literature regarding how top–down forecasts might improve on individual forecasts, especially the effect of correlation on the top–down approach. In the literature, there were arguments about whether positive or negative correlation would benefit the top–down approach. We conducted simulation experiments, assuming series share a common model and common seasonality within a group, to quantify the effect of correlation on the individual and grouping approaches in terms of forecasting accuracy. Our simulation results reveal that, when there are

two items in the group, the individual approach outperforms the grouping approach only when the correlation is very strongly positive. The grouping approach is better than the individual approach most of the time, with the benefit greater when correlation is negative. When there are more than two items in the group, the individual approach never outperforms the grouping approach in our simulations. DGSI is better for lower $|P^+/P^-|$ and WGSI is better for higher $|P^+/P^-|$.

Our current models do not take into account trend components. However, we have demonstrated that, for the additive model, the correlation between demands comes from the random error terms, with or without trend.

The conclusions from this chapter are general. Further research can build on the results and insights offered by this chapter and investigate the effect of correlation between demands by examining different models and assumptions.

REFERENCES

Brown, R. G. (1959). *Statistical forecasting for inventory control*. New York: McGraw-Hill.

Chatfield, C. (2004). *The analysis of time series* (6th ed.). London: Chapman & Hall/CRC.

Chen, H., & Boylan, J. E. (2007). Use of individual and group seasonal indices in subaggregate demand forecasting. *Journal of the Operational Research Society*, *58*, 1660–1671.

Dalhart, G. (1974). Class seasonality – A new approach. *American Production and Inventory Control Society Conference Proceedings*. Reprinted in *Forecasting*, 2nd ed., APICS, Washington, DC (pp. 11–16).

Dekker, M., van Donselaar, K., & Ouwehand, P. (2004). How to use aggregation and combined forecasting to improve seasonal demand forecasts. *International Journal of Production Economics*, *90*, 151–167.

Duncan, G., Gorr, W., & Szczypula, J. (1993). Bayesian forecasting for seemingly unrelated time series: Application to local government revenue forecasting. *Management Science*, *39*, 275–293.

Duncan, G., Gorr, W., & Szczypula. J. (1998). *Forecasting analogous time series*. Working Paper 1998-4. Heinz School, Carnegie Mellon University.

Fildes, R., & Beard, C. (1992). Forecasting systems for production and inventory control. *International Journal of Operations and Production Management*, *12*(5), 4–27.

Fogarty, D. W., & Hoffmann, T. R. (1983). *Production and inventory management*. Cincinnati, OH: Southwestern Publishing Co.

Lin, S. P., & Bendel, R. B. (1985). Algorithm AS213: Generation of population correlation matrices with specified eigenvalues. *Applied Statistics*, *34*, 193–198.

McLeavey, D. W., & Narasimhan, S. L. (1985). *Production planning and inventory control*. Boston, MA: Allyn and Bacon, Inc.

Muir, J. W. (1983). Problems in sales forecasting needing pragmatic solutions. *APICS Conference Proceedings* (pp. 4–7).

Ouwehand, P. (2006). *Forecasting with group seasonality*. Unpublished PhD thesis, Technische Universiteit Eindhoven, The Netherlands.

Schwarzkopf, A. B., Tersine, R. J., & Morris, J. S. (1988). Top–down versus bottom-up forecasting strategies. *International Journal of Production Research*, *26*, 1833–1843.

Shlifer, E., & Wolff, R. W. (1979). Aggregation and proration in forecasting. *Management Science*, *25*, 594–603.

Withycombe, R. (1989). Forecasting with combined seasonal indices. *International Journal of Forecasting*, *5*, 547–552.

Xu, K., & Evers, P. T. (2003). Managing single echelon inventories through demand aggregation and the feasibility of a correlation matrix. *Computers and Operations Research*, *30*, 297–308.

APPENDIX A. SIMULATION DESIGNS

Two Series

Quarterly seasonality was assumed in the simulations with four different seasonal profiles as shown in Tables A1 and A2.

Table A1. Seasonal Profiles for the Additive Model.

	Q1	Q2	Q3	Q4
No seasonality (NS)	0	0	0	0
Weak seasonality (WS)	−5	−10	5	10
Low, low, low, high (LLLH)	−20	−15	−15	50
Low, high, low, high (LHLH)	−25	25	−25	25

Table A2. Seasonal Profiles for the Mixed Model.

	Q1	Q2	Q3	Q4
No seasonality (NS)	1	1	1	1
Weak seasonality (WS)	0.9	0.8	1.1	1.2
Low, low , low, high (LLLH)	0.6	0.7	0.7	2
Low, high, low, high (LHLH)	0.5	1.5	0.5	1.5

The aim is not to attain comprehensiveness of seasonal profiles, but to choose a few commonly occurring profile shapes to check whether they affect the rules. WS represents a weak seasonality and LLLH represents a situation where there is a single very high season (e.g., in the final quarter of the year, with higher demand before Christmas). LHLH represents alternative low and high seasons.

The underlying mean for one item is fixed to be 50, and the mean of the other item in the group varies. It can take a value of 50, 100, 200, 300, 400, 500, 5,000 or 50,000, representing a ratio of 1, 2, 4, 6, 8, 10, 100 or 1,000.

Variances of the random error terms in the models are generated using power laws of the form $\sigma^2 = \alpha\mu^\beta$, where μ is the underlying mean, and α and β are constants (Brown, 1959). Our preliminary results agreed with Shlifer and Wolff (1979) that the α parameter does not affect the rules because it appears on both sides of the rule and can be cancelled out. Therefore, only the β parameter is allowed to vary in these power laws. We choose α to be 0.5 and β to be 1.2, 1.4, 1.6 or 1.8.

Variances of series of a group may follow power laws, but different series in a group may not follow the same power law. Therefore, we also simulate situations in which non-universal power laws are applied on a group. Series 1 in the group follows one law and series 2 follows the other law.

Series 1: $\sigma_i^2 = 0.75 \times 0.5\mu_i^{1.5}$
Series 2: $\sigma_i^2 = 1.25 \times 0.5\mu_i^{1.5}$

Alternatively, it may be assumed that the series follow no power laws. In this case, various combinations of mean and variance values have been identified, somewhat arbitrarily, for experimentation, as shown in Table A3.

Table A3. Arbitrary Variance Values.

Mean 1		50	50	50	50	50	50	50	50
Mean 2		50	100	200	300	400	500	5,000	50,000
No law									
Low	V1	100	100	100	100	100	100	100	100
Low	V2	100	225	1,600	2,500	3,600	4,900	62,500	1,562,500
Low	V1	100	100	100	100	100	100	100	100
High	V2	400	900	4900	8100	10000	22500	490000	49000000
High	V1	400	400	400	400	400	400	400	400
Low	V2	100	225	1600	2500	3600	4900	62500	1562500
High	V1	400	400	400	400	400	400	400	400
High	V2	400	900	4900	8100	10000	22500	490000	49000000

Data history is set to be 3, 5 or 7 years with the last year's observations used as the holdout sample. So the estimation periods are 2, 4 or 6 years.

The correlation coefficient is set to be –0.9, −0.6, −0.3, 0, 0.3, 0.6 and 0.9. This covers a wide range of correlation coefficients from highly negative to highly positive. These are correlations between the random variables in the model; they are also correlations between deseasonalised demands.

More than Two Series

We assume that the underlying mean values in a group follow a lognormal distribution. The details can be found in Table A4.

Table A4. Mean Values of the Lognormal Distribution.

Standard Ratio	Standard Deviation	Mean of the Logarithm	
		4	6
2	0.69	69	513
6	1.79	272	2009
10	2.30	774	5716
30	3.40	17749	131147

Each combination (2 means $\times$ 4 standard ratios) is replicated 50 times. MSE values are averaged over the 50 replications and then the results are compared. The purpose of replicating the lognormal distributions is to reduce randomness, especially when the group size is small (e.g., 4 items in the group) as the lognormal distribution may not be apparent. Such replication of distributions can also reduce the risk of some unusual values distorting the simulation results. For each replication of the lognormal distributions, 500 replications of the simulation are run. So, for each parameter setting, a total of 25,000 replications are run: 50 to replicate the lognormal distribution and 500 to replicate the estimation and forecasting process to reduce randomness (for each of the 50 distribution replications).

Variances are generated using only the universal power laws. The β parameter takes the values of 1.2, 1.4, 1.6 and 1.8. Non-universal power laws or arbitrary variance values are not examined in this chapter, owing to the greatly increased complexity of specifying the values.

APPENDIX B. SIMULATING CORRELATION MATRICES

Let

$$S = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}\sigma_i\sigma_j = \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{+}\sigma_i\sigma_j - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{-}\sigma_i\sigma_j \tag{B.1}$$

where $\rho_{ij}^{+} = \rho_{ij}$ if $\rho_{ij} > 0$ and $\rho_{ij}^{+} = 0$ otherwise. $\rho_{ij}^{-} = -\rho_{ij}$ if $\rho_{ij} < 0$ and $\rho_{ij}^{-} = 0$ otherwise.

$$s \leq \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{+}\sigma_{\max}^2 - \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{-}\sigma_{\min}^2 = \sigma_{\max}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{+} - \sigma_{\min}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{-} \tag{B.2}$$

By a similar argument, $s \geq \sigma_{\min}^2 \Sigma_{i=1}^{m-1} \Sigma_{j=i+1}^{m} \rho_{ij}^{+} - \sigma_{\max}^2 \Sigma_{i=1}^{m-1} \Sigma_{j=i+1}^{m} \rho_{ij}^{-}$, where $\sigma_{\min}^2$ is the minimum variance and $\sigma_{\max}^2$ the maximum variance.

Given all the σ values, it is clear that the sum of the positive correlation coefficients and the negative coefficients can be used to determine bounds on the cross-term corresponding to WGSI.

The same argument applies for DGSI. Let $s' = \Sigma_{i=1}^{m-1} \Sigma_{j=i+1}^{m} \rho_{ij}((\sigma_i/\mu_i)\,(\sigma_j/\mu_j))$,

$$\begin{aligned} CV_{\min}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{+} - CV_{\max}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{-} \leq s' \\ \leq CV_{\max}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{+} - CV_{\min}^2 \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \rho_{ij}^{-} \end{aligned} \tag{B.3}$$

where $CV_{\min}^2$ is the minimum coefficient of variation squared and $CV_{\max}^2$ the maximum coefficient of variation squared.

PART IV
OTHER APPLICATION AREAS OF FORECASTING

This page intentionally left blank

ECONOMETRIC COUNT DATA FORECASTING AND DATA MINING (CLUSTER ANALYSIS) APPLIED TO STOCHASTIC DEMAND IN TRUCKLOAD ROUTING

Virginia M. Miori

ABSTRACT

The challenge of truckload routing is increased in complexity by the introduction of stochastic demand. Typically, this demand is generalized to follow a Poisson distribution. In this chapter, we cluster the demand data using data mining techniques to establish the more acceptable distribution to predict demand. We then examine this stochastic truckload demand using an econometric discrete choice model known as a count data model. Using actual truckload demand data and data from the bureau of transportation statistics, we perform count data regressions. Two outcomes are produced from every regression run, the predicted demand between every origin and destination, and the likelihood that that demand will occur. The two allow us to generate an expected value forecast of truckload demand as input to a truckload routing formulation. The negative binomial distribution produces an improved forecast over the Poisson distribution.

Advances in Business and Management Forecasting, Volume 6, 191–216

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006012

1. INTRODUCTION

This chapter brings together the application areas of truckload routing and econometric count data analysis, whereas data mining provides a basis for selection of the statistical distribution underlying the count data regression. The combination provides the opportunity to solve the stochastic truckload routing problem (TRP) to optimality. Individually these areas have been researched; however, they have not been examined collectively. The literature presented therefore represents distinct areas of research.

The triplet model formulation was presented by Miori (2006) as an alternative to allow solution of the TRP through the means of dynamic programming. This formulation examined demand in combination with any necessary empty movements that would be incurred. An alternative representation of stochastic demand is easily facilitated through this reformulation.

Arunapuram, Mathur, and Solow (2003) noted that truckload carriers were faced with a difficult problem. They present a new branch and bound algorithm for solving an integer programming formulation of this vehicle routing problem with full truckloads (VRPFL). Time window and waiting cost constraints were also represented in the problem formulation. The resulting efficiency of the method was due to a column generation scheme that exploited the special structure of the problem. The column generation was used to solve the linear relaxation problems that arose at the nodes. The objective in solving the VRPFL was to find feasible routes that minimized cost. They noted that minimizing the total cost was equivalent to minimizing the cost of traveling empty.

Chu (2004) presented a heuristic that applied to both the VRP and the TRP (less-than-truckload versus truckload applications). The focus of the paper was a heuristic technique, based on a mathematical model, used for solution generation for a private fleet of vehicles. Outside carriers that provided less-than-truckload (LTL) service could be employed in the solution. In this paper, Chu addressed the problem of routing a fixed number of trucks with limited capacity from a depot to customers. The objective was the minimization of total cost.

The TL-LTL heuristic developed in this paper was designed not only to build routes, but also to select the appropriate mode. The first step was the selection of customers to be served by mode. LTL customers were stripped out and routes built using the classic Clarke and Wright (1964) algorithm. The final step was swapping customers between and within routes. Though this paper addressed the TRP, its primary focus was on the

LTL mode and in fact replicated a typical "cluster first-route second" approach.

Gronalt, Hartl, and Reimann (2003) used a savings-based approach to the TRP with time window constraints. The objective was to minimize empty vehicle movements that generate no revenue. They provided an exact formulation of the problem and calculated a lower bound on the solution based on a relaxed formulation using network flows. They further presented four different savings-based heuristics for the problem. There were generally two different problem classes. The first was concerned with dynamic carrier allocation and the second class of problems dealt with the pickup and delivery of orders by a fleet of vehicles. This problem belonged to the second class.

Gronalt et al. strived for the goal of minimizing empty vehicle movements for the truckload portion of the shipments. The results were extended to a multidepot problem as well. The algorithms included the savings algorithm, the opportunity savings algorithm that incorporated opportunity costs into calculation of savings values, and the simultaneous savings approach that accepted a number of savings combinations at each iteration that satisfy a given condition.

The TRP has also been addressed in the literature as a multivehicle truckload pickup and delivery problem. Yang, Jaillet, and Mahmassani (2000) presented a problem in which every job's arrival time, duration, and deadline are known. They considered the most general cost structure and employed alternate linear coefficients to represent different cost emphases on trucks' empty travel distances, jobs with delayed completion times, and the rejection of jobs. Note again that this problem did not require all potential loads to be serviced. The problem was presented first as an off-line problem with cost minimization as its objective. A rolling horizon and real-time policies were then introduced.

The authors examined a specific stochastic and dynamic vehicle routing problem. They devised heuristic stationary policies to be used by the dispatcher for each situation, at each decision epoch. The policies were based on observations of the off-line problem and were intended to preempt empty movements. If a vehicle was already en route to a job, it could be dispatched to serve another job based on these rules. A series of simulations showed that the policies developed were very efficient.

Cameron and Trivedi (1986) discussed count data analysis in the premier issue of the *Journal of Applied Econometrics*. This work examined Poisson, compound Poisson, negative binomial, and more generalized models of count data. Their work continued and culminated in the publication of a

seminal text in the area (Cameron & Trivedi, 1998). Greene (2003) and Agresti (2002) both presented and applied count data models in text format.

Data mining, specifically clustering of data, has been used in areas extending from quality analysis, traffic safety to demand forecasting. BaFail (2004) forecasted demand of airline travelers in Saudi Arabia. Neural networks were applied to 10 years of demand data. Bansal, Vadhavkar, and Gupta (1998) used traditional statistical techniques to determine the best neural network method to use for inventory forecasting in a large medical distribution company.

The area of count data models applied to stochastic demand in the freight transportation industry has been the subject of little or no research. As such, the literature presented here represents alternate application areas for count data models in an effort to draw parallels between the application areas.

Arunapuram et al. (2003) treated the demand deterministically, thus requiring no distribution assumptions. Powell (1996) considered the stochastic demand associated with truckload transportation and presented a multistage model to address it. He did not however offer a specification of distribution of demand. Frantzaskakis and Powell (1990) assumed a Poisson distribution of demand with a mean equal to the historical frequency.

Recreational travel was analyzed according to household factors, typical frequency of travel and ethnicity by Bowker and Leeworthy (1998). Haab (2003) examined recreational travel demand with consideration of temporal correlation between trips. Hellstrom (2006) studied tourism demand in Sweden again according to number of leisure trips and number of overnight stays using count data. Jang (2005) used count data modeling to generate trips, based on socioeconomic characteristics, as a way of overcoming the inherent shortcomings of linear regression. He first applied the Poisson regression model only to find overdispersion, thus turning to the negative binomial model as the ultimate representation. Ledesma, Navarro, and Perez-Rodriguez (2005) used count data modeling to examine the hypothesis that repeated visits to the same location was actually the result of asymmetrical information. A left-truncated Poisson regression was applied.

Chakraborty and Keith (2000) used truncated count data travel cost demand models to estimate the demand for and value of mountain biking in Moab, Utah. Both truncated Poisson and truncated negative binomial distributions were used to determine that mountain biking had a higher value to the area than other recreational activities.

Puig (2003) presented the aspects of the Poisson distribution that make it the most widely used distribution in count data analysis: it is closed under addition and the sample mean is actually the maximum likelihood estimator

(MLE) of the mean of the distribution. Other discrete distributions with two parameters were discussed in an effort to find these same aspects and thus provide an alternate distribution for count data analysis. The general Hermite distribution was discussed at great length though it is only partially closed under addition. The Poisson distribution is a subfamily of the general Hermite distribution. Puig and Valero (2006) carried this discussion farther and characterized all two-parameter discrete distributions that are also partially closed under addition. They limited consideration to only those distributions for which the sample mean was the same as the MLE of the mean. The count data models included the negative binomial, Neyman A, Polya Aeppli, Poisson-inverse Gaussian, Thomas, and general Hermite. The authors also presented a theorem that allows extension to other two-parameter discrete distributions for use in count data models.

Negative binomial regressions have also been employed in assessing intersection-accident frequencies. Poch and Mannering (1996) looked at intersections in Bellevue, Washington. Rather than using the count data model with a negative binomial distribution underlying the regression, they used a maximum likelihood estimation procedure. Karlaftis and Tarko (1998) discussed improved accident models using panel data. Rather than employing the Poisson distribution, which has been the source of criticism, they employed clustering of the data and subsequently applied multiple negative binomial distributions to represent the accident frequency data.

2. TRP TRIPLET FORMULATION

The most basic assumption of the TRP is that freight is delivered by truck, in full truckload quantities. This means that a full truckload quantity is defined as a single unit of demand and may not be split between multiple vehicles.

The formulation utilizes a triplet concept rather than a lane concept (Miori, 2006). A node triplet is composed of two origin/destination pairs in succession in which the first destination location is the same as the second origin location. It may be made up of two loaded movements, one empty movement and one loaded movement or two empty movements.

The TRP typically results in a loaded lane (origin/destination pair) movement followed by an empty lane movement. The movements are tied together and a logical connection between the associated loaded and empty movements is created. The cost of the triplet is therefore the sum of the costs

of the individual movements. The transit and work times are the sums of the transit and work times of the individual movements.

Traditionally, a route is composed of a series of lanes to be serviced by a single vehicle. We extend the notion of a route to consider a series of triplets. We must begin and end a route at the same node, known as a domicile. A route is therefore composed of a sequence of triplets. Within any route that's generated, the final node of a triplet must match the subsequent triplet origin (first node). This organization of these locations has been chosen to consider the opportunity cost of each movement.

Time conditions are placed on the handling of freight by the customers. Loads must be picked up and delivered on particular days of the week during specified times of the day. Some customers will be more restrictive than others in specifying their time windows.

The remaining conditions are imposed by the Department of Transportation (DOT). Restrictions are placed on the number of driving hours in a given week and number of work hours in a given week.

A primary advantage of the triplet formulation is a more natural treatment of variation in length and cost of empty movements. Stochastic demand is however still present.

The first terms in the TRP notation require no subscripts. They address parameters set by the data itself and the carrier preference within DOT hours of service restrictions for number of hours allowed on a tour.

N The number of nodes.
V The number of vehicles.
H Total allowed transit hours per route.

The time window specifications are particular to the location of each node. These require the subscript i, where $i = 1, \ldots, N$.

D_i Departure time from node.
$[e_i,l_i]$ Early and late arrival times for node i.

The demand for service and the travel time are provided for each lane pair. Lanes are designated as ij, where $i = 1, \ldots, N$ and $j = 1, \ldots, N$.

y_{ij} 1 if lane ij represents loaded movement.
0 otherwise.
t_{ij} Travel time between node i and j.

The costs per triplet per vehicle and the decision variables require three subscripts, i, j, and k, where $i = 1, \ldots, N$; $j = 1, \ldots, N$; and $k = 1, \ldots, N$.

c_{ikj} Cost to service triplet ijk.
x_{ikj} 1 if triplet ijk served by vehicle v.
0 otherwise.

The objective may be stated as a cost minimization or a profit maximization function. The financial structure of the transportation provider will dictate which of these approaches is preferred. It is likely that a private fleet would choose cost minimization, whereas a for-hire carrier would select profit maximization. In this chapter, the formulation is presented as a cost minimization problem.

$$\min \sum_{ijk} c_{ijk} x^{v}_{ijk} \tag{1}$$

A routing problem always has specific characteristics that must be modeled, and as such, the constraints may be easily categorized into sets. The first set of constraints is for load satisfaction. It guarantees that each lane pair with demand (available load) is served at least once. A lane may be served as the first leg of a triplet (origin node to intermediate node) or as the second leg of the triplet (intermediate node to destination node).

$$\sum_{vk} y_{ij}(x^{v}_{ijk} + x^{v}_{kij}) \geq 1 \quad \forall y_{ij} = 1 \tag{2}$$

The load must be carried and additional empty movements may also utilize the same lane. If a lane has no demand, it need not be used in the overall solution. The nature of the optimization ensures that lanes with no demand would only be used or reused for empty movements in support of an optimal solution. Since we employ a triplet formulation, we must combine lane level information (demand) with triplet level information (decision variables representing inclusion or exclusion of a triplet in the solution) in this constraint set.

The next set of constraints preserves traditional conservation of flow and schedule feasibility. The conservation of flow constraints ensure that every vehicle departs from the final node of every triplet it serves.

$$\sum_{ijlmv} (x^{v}_{ijk} - x^{v}_{klm}) = 0 \quad \forall k \tag{3}$$

The schedule feasibility constraints ensure a logical progression through the routes (a vehicle may not depart before its arrival), and forces adherence to time constraints (time windows are satisfied for the pickup and delivery points as well as the domicile). Because time windows address individual locations, the time window constraints reflect the individual nodes and not the entire triplet. These are standard constraints used in the TRP.

$$D_i + t_{ij} \leq D_j \quad \forall i, j \tag{4}$$

$$e_i \leq D_i \leq l_i \quad \forall i \tag{5}$$

The remaining constraint set ensures that the routes satisfy the DOT regulations for single drivers. Each route is restricted to a maximum number of hours in transit. The DOT recently revised these regulations, but the new regulations have been challenged. The model discussed in this chapter allows for flexibility in the statement of these constraints to reflect this uncertainty.

$$\sum_{ijk} (t_{ij} + t_{jk})\, x_{ijk}^{v} \leq H \quad \forall v \tag{6}$$

The decision variables must take on binary values.

$$x_{ijk}^{v} \in \{0, 1\} \tag{7}$$

3. DATA PREPARATION

The need for a large volume of data that accurately represents national freight volumes arises from the use of data mining cluster analysis and the econometric count data model. The cluster analysis revealed only four fields in the load-specific data that contributed to the cluster structures. The count data model, however, combines the use of national freight flow information with load-specific data that represents the truckload freight operation under consideration. The national freight data was gathered from the Bureau of Transportation Statistics (BTS), whereas the carrier-specific data was collected from truckload freight operations with domiciles in 14 different cities. We first discuss the BTS data.

3.1. Bureau of Transportation Statistics Data

Within the BTS, freight surveys are completed approximately every five years. This cross-sectional data is quite extensive and covers freight

volumes, hazardous material issues, safety, and limitations specific to mode. The Commodity Flow Surveys within Intermodal Transportation Database and the Highway Data Library contain numerous tables applicable to truckload transportation. This data came specifically from two tables that contain outbound and inbound shipment characteristics.

The BTS eliminated data records with anomalies and errors before publishing its data. Therefore, an imbalance exists in the freight volume reported. The overall freight supply exceeds the freight demand. Before completing any analysis that relies on balance of freight, the inequity must be resolved. The freight supply was scaled to eliminate the appearance of freight that traveled the country indefinitely.

The selected tables contain data that reflect specific origins and destinations for freight. These origins and destinations include all major metropolitan areas within the 50 states as well as rural and less populated areas. Inbound and outbound freight volumes for each location within every state are reported.

The volumes of freight inbound to each location and the volumes of freight outbound from each location provide the first useful application of the BTS data. These values are used in the subsequent count data model as explanatory variables. In addition, the data is used to generate indices for supply/demand centers at each metropolitan area. Note that these accumulations of data were all produced using the scaled supply figures and the reported demand figures.

The supply/demand index points to the nature of the particular metropolitan or rural area. A strict demand center is one in which all freight is delivered to that location, but none is available for pick-up. It is designated by an index of 1.00. A strict supply center is just the opposite. It is one in which no freight is destined for that location, but freight is always available to be shipped. An index of -1.00 is designated for a supply center. A completely balanced location will have an index of 0.00. The supply/demand index may therefore fall anywhere in the interval from -1.00 to 1.00.

The BTS data represents 138 metropolitan areas or regions in the United States, including Alaska and Hawaii ($i = 1, \ldots, 138$). The supply/demand index is determined using the following formula:

$$\text{index}_i = \frac{\text{demand}_i\ (\text{tons}) - \text{scaled supply}_i\ (\text{tons})}{\text{demand}_i\ (\text{tons}) + \text{scaled supply}_i\ (\text{tons})} \tag{8}$$

The final supply/demand indices were cross-referenced with the load-specific data to be discussed next.

3.2. Load-Specific Truckload Data

The second data source contained information on a load-by-load basis. The original database contained over 34,000 records that included date, origin, volume, number of truckloads, and mileage traveled for each load. There were 15 origin terminal locations. There were no specific destination names provided in the data, but there were nongeographic destination designations. Using mileage, origin locations, and overall demand patterns, we were able to triangulate 50 destinations.

The discernment of the destination then allowed the preparation of data for the cluster analysis and the count data model to continue. Each origin and destination was matched with its supply/demand index. In addition, the outbound volume for each origin and the inbound volumes for each destination were also matched.

On the basis of date, each record was assigned additional variables which reflected the day of week and the quarter of the year in which the load was carried. These variables were further used to create a series of dummy variables as required to model the timeframe. Note that each of the time window dummy variables corresponds directly to the level of the decision tree in which a load might be serviced.

The chosen aggregation method accumulated loads by origin, destination, and date. The minimum value for demand is a single truckload. To match demand data to either the Poisson or negative binomial distributions, we performed a linear transformation of the data by decrementing the number of loads by one (Fig. 1). The mean and variance of the decremented data do not appear to be the same resulting in a violation of that property of the Poisson distribution. This difference between the mean and variance leads to the possibility of overdispersion in the data.

The issue of unequal mean and variance drives us to consider the negative binomial as a basis for the count data model. We cannot make a complete assessment of the equality from the descriptive statistics. We can however test for the equality as we apply both the Poisson and negative binomial regressions. Further, by clustering our data using data mining, we are able to examine the equal mean and variance property in detail.

4. DATA MINING: CLUSTERING METHODS

Research conducted on the TRP with stochastic demand has typically generalized demand to a Poisson distribution. Rather than making this

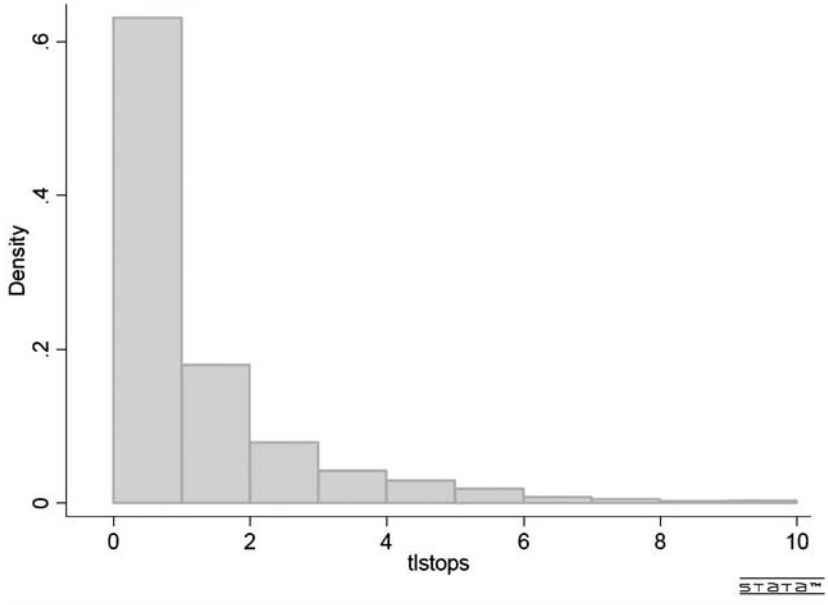

Fig. 1. Demand Frequency Histogram.

assumption, we performed a series of cluster analyses using SPSS Clementine Data Mining Software. Chi square goodness-of-fit tests were used to examine the fit of each distribution to the data.

The demand data was imported for use in Clementine. Once there, a series of cluster analyses were performed using both two-step and K-means methods. The two-step method, just as most data mining techniques does not require the user to know what groups might exist before performing the analysis. Rather than trying to predict outcomes, it attempts to uncover patterns within the data. The first step is a single path through the data in which a manageable set of subclusters are created. In the second step, a hierarchical method is employed which progressively merges subclusters into larger clusters. Since the number of clusters is not specified ahead of time, the second step concludes when combining clusters can no longer result in effective identification of data characteristics.

K-means clustering also begins with no requirement to know what groups might exist. It does, however, ask the user to specify the number of clusters to be generated. A set of starting clusters is derived from the data. The data records are then assigned to the most appropriate cluster. The cluster centers

are updated to reflect all new records assigned. Records are again checked to determine whether they should be reassigned to a different cluster. The process continues until the maximum number of iterations is reached or the change between iterations does not exceed a specified threshold.

Fig. 2 shows the cluster characteristics for the two-step method. Figs. 3–5 all show the characteristics for the K-means method, first with three clusters, then four clusters, and finally five clusters. The quality of each clustering results was examined and tested using the chi square goodness-of-fit test. The null hypotheses state that the data within each cluster follow a Poisson distribution or a negative binomial distribution.

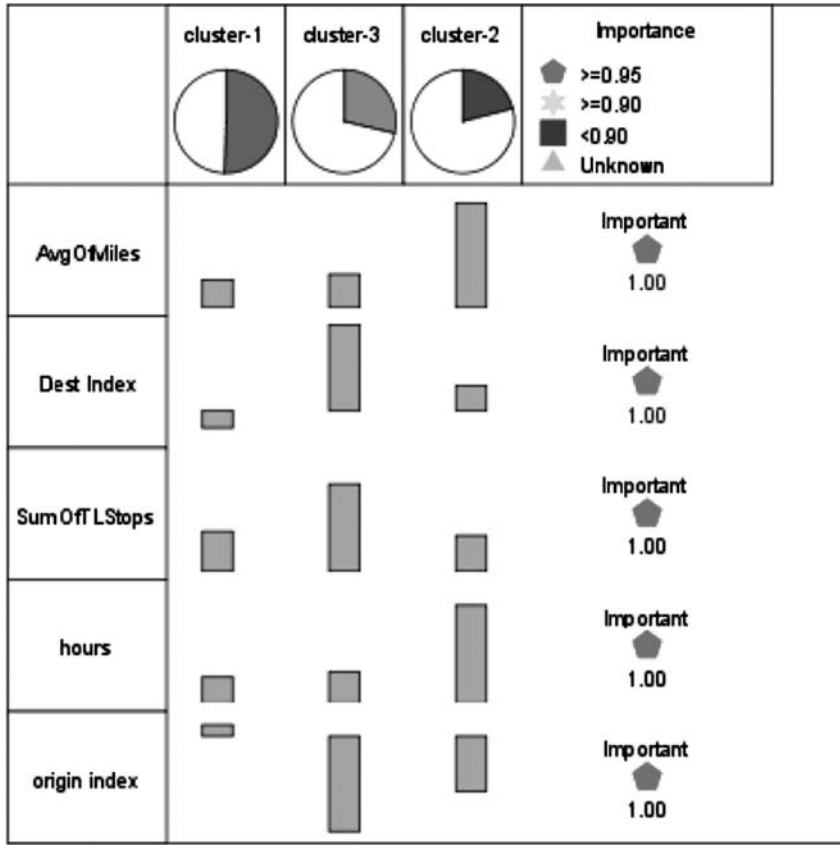

Fig. 2. Two-Step Cluster Characteristics.

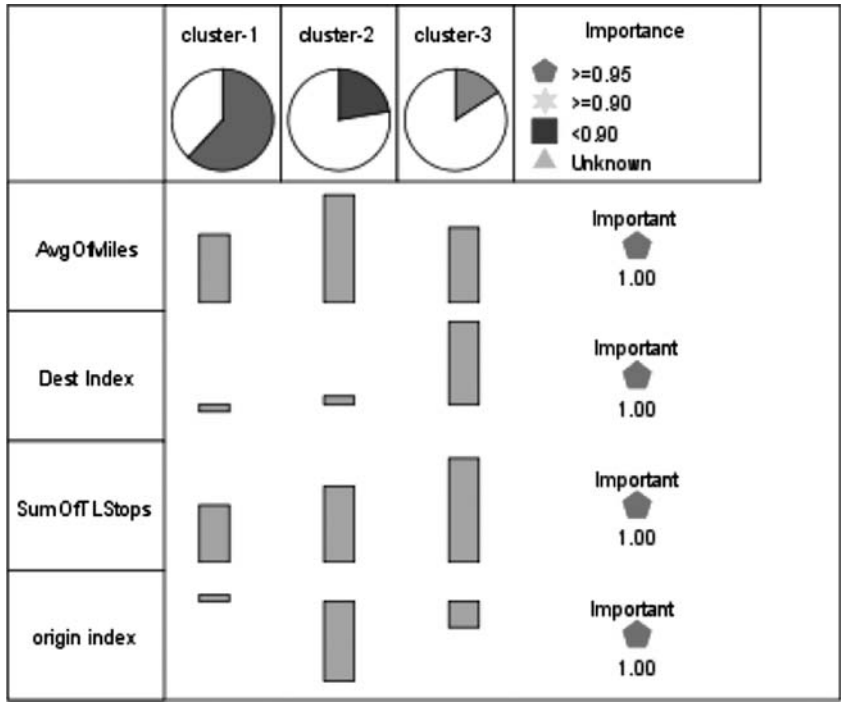

Fig. 3. K-Mean 3 Cluster Characteristics.

The results are presented in Table 1. The preferred clustering method was in fact the two-step method with chi square values of 20.48 for the first cluster, 16.58 for the second cluster, and 51.70 for the third cluster. Though these values do not provide certainty that the negative binomial is the ideal distribution to apply, they certainly do argue that the negative binomial is a far superior fit for the data than the Poisson. In addition, the results argue that the two-step clustering method has produced the superior clustering of the data.

The clusters were fit to a negative binomial distribution and that supports our initial premise that the Poisson distribution was not the ideal distribution to underlie the count data regression analysis. The specific characteristics of the clusters in this approach are presented in Table 2.

The count data analysis still carries forward with both the Poisson and negative binomial distributions. The continued analysis of both distributions provided even greater support of the negative binomial as the preferred distribution to represent stochastic truckload demand.

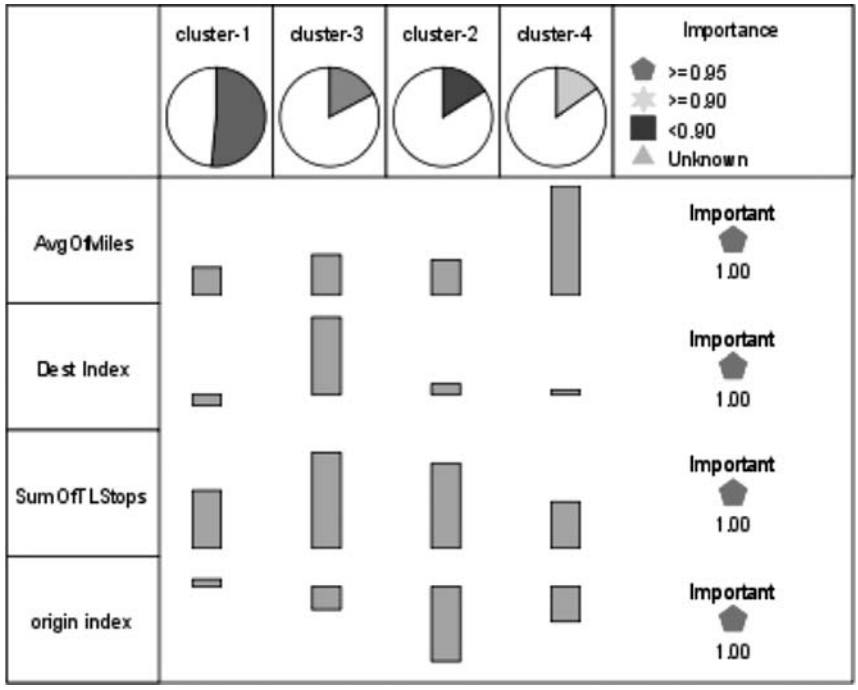

Fig. 4. K-Mean 4 Cluster Characteristics.

5. DISCRETE CHOICE MODELS

Count data models are discrete choice models. The response (dependent) variable takes on discrete values within a range of possibilities. In the case of the TL data, these discrete values represent the number of loads available between a particular origin and destination on a specific day of the week y_{ijk}. The range of values is [0,15].

The probability that the response variable takes on particular values is written

$$\Pr(Y_{ijk} = y_{ijk} | x_{ij}) \tag{9}$$

and is calculated using the probability mass function for the associated probability distribution.

The predominant probability distribution that underlies the count data model is the Poisson distribution with λ_{ijk}, the expected number of loads per

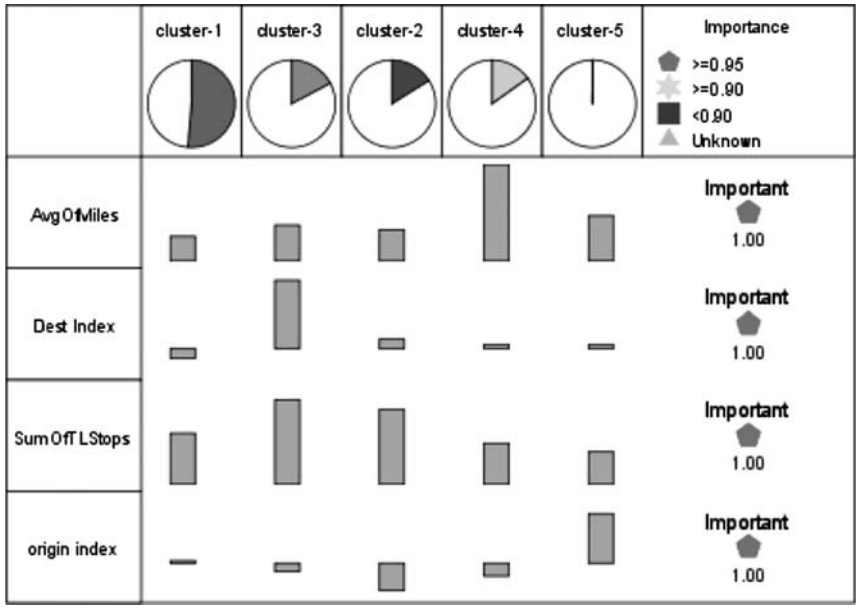

Fig. 5. K-5 Cluster.

Table 1. Chi Square Goodness-of-Fit Test Results.

Cluster Designation	Geometric Distribution	Negative Binomial Distribution	Poisson Distribution	Degrees of Freedom	Cluster size
TS C1	20.48	20.48	85.35	3.00	2,076.00
TS C2	16.58	16.58	75.38	3.00	860.00
TS C3	56.21	51.70	6,371.98	10.00	1,180.00
KC5 C1	134.05	142.75	561.93	7.00	2,113.00
KC5 C2	61.79	61.79	1,155.15	9.00	663.00
KC5 C3	26.04	21.33	2,180.85	8.00	710.00
KC5 C4	19.70	19.70	73.95	3.00	623.00
KC5 C5	0.37	0.00	0.74	0.00	7.00
KC4 C1	136.11	144.11	564.82	7.00	2,119.00
KC4 C2	61.79	61.79	1,155.15	8.00	663.00
KC4 C3	26.04	21.33	2,180.85	8.00	710.00
KC4 C4	19.38	19.38	73.07	3.00	624.00
KC3 C1	119.86	122.37	679.55	7.00	2,545.00
KC3 C2	115.13	115.13	974.18	8.00	920.00
KC3 C3	17.92	17.92	2,383.34	9.00	651.00

Table 2. Two-Step Cluster Characteristics.

	Cluster 1 Mean	Cluster 2 Mean	Cluster 3 Mean
Miles	339.59	1,273.967	403.037
Destination Index	−0.03	0.044	0.149
Hours	6.792	25.479	8.075
Origin Index	0.014	−0.067	−0.116
Demand	1.385	1.235	3.031

day, as the response variable. Each value of the response variable is considered to have been drawn from a Poisson distribution with the mean of λ_{ijk}. When the Poisson distribution is used, the property that mean and variance of the distribution are equal applies.

The Poisson regression may not always provide the best fit for data being modeled. The Poisson model has the property of equal mean and variance of the data. If instead, the variance exceeds the mean, we conclude that the data is overdispersed. If the variance is exceeded by the mean, we conclude that the data is underdispersed. The risk of overdispersion is significant. A comparison of the sample mean and the sample variance can provide a strong indication of the magnitude of the overdispersion.

The Poisson regression is in fact nested within an alternate distribution applied to count data models, the negative binomial distribution. The negative binomial provides greater flexibility in modeling the variance than the Poisson. Owing to this nesting, we can perform a hypothesis test to determine whether the Poisson distribution is appropriate.

5.1. Independent Variables

Historic demand patterns may be broken down into base demand and seasonal fluctuations. The base demand allows lanes (origin/destination pairs) to be compared to each other and even placed on a ratio scale. Seasonal or cyclical fluctuations will be considered uniquely for each lane. In this analysis, we will first consider two indicators of base demand, the demand for freight leaving the load origin (*origloadsout*) and the demand for freight entering the load destination (*destloadsin*). We also include a seasonal adjustment that necessitates the use of seasonal dummy variables (*seasonQ2*, *seasonQ3*, *seasonQ4*). When using dummy variables, we avoid collinearity by specifying one fewer variable than circumstances to be

represented. Therefore, dummy variables representing Q2, Q3, and Q4 are included in the model, whereas Q1 is considered the base timeframe.

We turn our attention next to the characteristics of the origins (*origctr*) and destinations (*destctr*) for each load. Every location considered in a routing scenario may be viewed at the extremes as a supply center or a demand center. We designate a strict demand center (no available loads to depart from the location) with a 1.00. A strict supply center (no available loads destined for the location) with a -1.00. A 0.00 represents a balanced location with equal availability of loads destined for and departing from that location. Most locations will fall somewhere else along the continuum from -1.00 to 1.00.

The distance traveled to deliver each load (*dist*) is taken into consideration as well. A number of issues arise from the inclusion of very long hauls. This portion of the analysis, however, is concerned only with the predictive nature of distance. Rather than incur a very long and costly movement, many companies will try to source the load from an alternate location. We therefore anticipate a decline in number of loads as distance increases. Owing to the overhead of truckload travel, there also tend to be fewer loads carried short distances. (LTL compares more favorably for short distances.) The nature of truckload transportation is that there is a lower frequency associated with loads traveling very short or very long distances. For this reason, we will also include a variable for squared (*distsq*) distance and cubed distance (*distcb*).

The time windows that exist in the TL mode are typically specified on a day-by-day basis and represent when the load must depart from its origin rather than when it must arrive at its destination. Receipt of freight occurs primarily on weekdays, but may have occasion to occur on weekends as well. Therefore, we consider all days of the week. On the basis of this, time windows may be modeled using two dummy variables for the three time windows of the week: Monday–Tuesday, Wednesday–Thursday, and Friday–Saturday

(TimeWin2, TimeWin3). Time windows are the single factor that will result in variation in the probabilities generated by the count data model.

We now integrate both the dependent and independent variables into the log linear model specification of λ_{ijk}. It takes the following form:

$$\begin{aligned}\ln \lambda_{ijk} = {} & \beta_0 + \beta_1 \text{TimeWin2} + \beta_2 \text{TimeWin3} + \beta_3 \text{dist} + \beta_4 \text{origctr} \\ & + \beta_5 \text{destctr} + \beta_6 \text{origloadsout} + \beta_7 \text{destloadsin} + \beta_8 \text{seasonQ2} \\ & + \beta_9 \text{seasonQ3} + \beta_{10} \text{seasonQ4} + \beta_{11} \text{distsq} + \beta_{12} \text{distcb}\end{aligned} \tag{10}$$

6. REGRESSION DATA ANALYSIS

The aggregate truckload data combined with the indices produced from the BTS data provide the base data for the analysis. Before the completion of the regression, we generate and examine descriptive statistics on our fields to initiate a better understanding of the data.

Descriptive statistics for the truckload demand over all of the data must be generated as well as descriptive statistics for the truckload demand for load origin and destination pairs within the aggregated data. The descriptive statistics provide the first clue as to which type of regression may be more effective as the underlying distribution in the count data model.

6.1. Use of Stata

Stata Version 8 is the software product used in the count data model and the associated analysis. Stata is a statistical package used heavily in epidemiological and econometric analysis. It provides most capabilities necessary for data analysis, data management, and graphics. The count data model may be evaluated using the Poisson distribution or either of the negative binomial distributions.

6.2. Analysis Procedure

On examination of the overall histogram in Fig. 1, it became a possibility that the Poisson distribution might not be suitable for the count data model due to the high concentration of zero observations. Recall that we have decremented the demand by one in order to allow for analysis using a Poisson or negative binomial regression. We cannot make a definitive conclusion based solely on the histogram; therefore, we begin to supplement our knowledge through the generation of descriptive statistics. These descriptive statistics are presented in Appendix A. If the mean and variance of the truckload demand in the data appear to differ, one is directed toward the negative binomial distribution as a base for the count data model: if they appear consistent, the Poisson regression is indicated.

We have already discussed the examination of aggregate data as well as examination of data grouped by origin/destination pair. Both approaches provide valuable information. The success of one approach beyond the other will be a reflection of the nature of the specific data used in any

analysis. Running all data together in aggregate form allows the evaluation of the significance of the independent variables in the regression. Looking at individual origin/destination pairs relies on our acceptance of an assumption of conditional independence, but allows the focus to fall entirely on the time windows and seasonality associated with each load.

Ultimately, both approaches offer insight into the data and both approaches must be used in order to determine how to best model the data. Once regression coefficients have been established, the predicted counts may be generated as well as the mass functions that provide the probability that a particular number of loads occurs.

6.3. Count Data Model Results

Before completing the count data analysis, we must generate the supply and demand indices from the BTS data. Overall, the freight flows in the United States are really quite balanced. On calculation and examination of the supply/demand indices, we discover that over 60 percent of the locations contained in the database have supply/demand indices that fall between -0.10 and 0.10. Less than 7 percent of these locations have indices that fall below -0.50 or above 0.50.

All other data required for the count data regression was pulled directly from the lane-specific truckload database or from the general BTS summary. The field of greatest interest is truckload demand. We again examine the demand histogram, Fig. 1, providing initial justification for consideration of the Poisson or negative binomial regression.

The histogram provided evidence and a starting point. The analysis continued with the calculation of descriptive statistics. It provided indications that the Poisson regression may not be suitable to represent the truckload demand data. Recall that the primary reason to discard the Poisson regression was overdispersion. The difference between the mean and the variance of this data appeared to be significant enough to point us in the direction of the negative binomial.

To test the hypothesis and make the determination of the appropriate distribution, the first run performed in Stata was the Poisson regression on all independent variables. The output from this series of runs is found in Appendix B. Several of the variables showed p values that indicated a lack of significance. More importantly, however, the incredibly high chi squared statistic of over 1,379, and its associated p value of 0 indicated that we reject

the null hypothesis and conclude that the data did not follow a Poisson distribution. We indeed appeared to suffer from overdispersion.

The appropriate follow-up analysis was to perform a negative binomial regression using the NB2 model on that same data. There are other more complicated distributions that serve as bases for the count data regression analysis, but the burden of application is significant and the results do not have direct interpretations.

The preliminary run of the NB2 regression again included all of the independent variables. The dummy variables for Quarter4, Window2, and Window3 and the distance variable were found to be insignificant. In the ensuing run, the insignificant variables were eliminated but the distance variable remained. Recall that the regression included distance, and higher order terms of the distance in anticipation of a nonlinear relationship between distance and demand. Therefore, the distance variable remained in the regression despite its apparent insignificance.

In the remaining count data run, the negative binomial NB2 model was again evaluated. All of the independent variables remained significant with the exception of the distance variable. The next step was a hypothesis test of the Poisson distribution. The null hypothesis stated that $\alpha = 0$, indicating that the underlying distribution was indeed a Poisson distribution. The value of α generated was 1.5105. The p-value of the likelihood ratio test of $\alpha = 0$ was equal to zero. Therefore, we rejected the null hypothesis and assumed that the underlying distribution was not Poisson, but was negative binomial.

The negative binomial count data regression was then used as the base for calculating expected counts and probabilities for every origin/destination pair within the data. The expected counts were easily generated within Stata. The probabilities, however, required the development of a C++ program. The probability mass function for the negative binomial distribution was embedded into this code and therefore used to generate the likelihood or probability associated with each count. The counts and probabilities may now be integrated into any continuing analysis.

7. CONCLUSIONS AND FUTURE RESEARCH

The use of data mining techniques and an econometric count data model to forecast for the TRP offers many advantages. More characteristics of the demand patterns are considered as well as direct consideration of time windows without inclusion of time window constraints in the TRP. It also

opens the door to the use of the forecast in a more sophisticated fashion, as an expected value. It facilitates the eventual solution of the TRP using dynamic programming.

In this research, cluster analysis provided strong supporting evidence that the Poisson distribution is unsuitable for representation of stochastic demand data. It also provided sufficient evidence to conclude that the negative binomial distribution provided a solid fit. A single regression was performed to generate the forecasted demand. We may extend this research to create individual count data regression models for each of the demand clusters.

REFERENCES

Agresti, A. (2002). *Categorical data analysis* (2nd ed.). Somerset, NJ: Wiley.

Arunapuram, S., Mathur, K., & Solow, D. (2003). Vehicle routing and scheduling with full truckloads. *Transportation Science*, *37*(2), 170–182.

BaFail, A. O. (2004). Applying data mining techniques to forecast number of airline passenger in Saudi Arabia (Domestic and international travels). *Journal of Air Transportation*, *9*(1), 100–115.

Bansal, K., Vadhavkar, S., & Gupta, A. (1998). Neural networks based forecasting techniques for inventory control applications. *Data Mining and Knowledge Discovery*, *2*, 97–102.

Bowker, J. M., & Leeworthy, V. R. (1998). Accounting for ethnicity in recreation demand: A flexible count data approach. *Journal of Leisure Research*, *30*(1), 64–78.

Cameron, A. C., & Trivedi, P. K. (1986). Econometric models based on count data: Comparisons and applications of some estimators and tests. *Journal of Applied Econometrics*, *1*, 29–53.

Cameron, A. C., & Trivedi, P. K. (1998). *Regression analysis of count data*. Cambridge, UK: Cambridge University Press.

Chakraborty, K., & Keith, J. E. (2000). Estimating the recreation demand and economic value of mountain biking in Moab, Utah: An application of count data models. *Journal of Environmental Planning and Management*, *43*(4), 461–469.

Chu, C.-W. (2004). A heuristic algorithm for the truckload and less-than-truckload problem. *European Journal of Operational Research*, 1–11.

Clarke, G., & Wright, J. W. (1964). Scheduling of vehicles from a central dept to a number of delivery points. *Operations Research*, *12*, 568–581.

Frantzaskakis, L. F., & Powell, W. B. (1990). A successive linear approximation procedure for stochastic, dynamic vehicle allocation problems. *Transportation Science*, *24*(1), 40–57.

Greene, W. H. (2003). *Econometric analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Gronalt, M., Hartl, R. F., & Reimann, M. (2003). New savings based algorithms for time constrained pickup and delivery of full truckloads. *European Journal of Operational Research*, *151*, 520–535.

Haab, T. C. (2003). Temporal correlation in recreation demand models with limited data. *Journal of Environmental Economics and Management*, *45*, 195–212.

Hellstrom, J. (2006). A bivariate count data model for household tourism demand. *Journal of Applied Econometrics*, *21*(2), 213–227.

Jang, T. Y. (2005). Count data models for trip generation. *Journal of Transportation Engineering*, *131*(6), 444–450.

Karlaftis, M. G., & Tarko, A. P. (1998). Heterogeneity considerations in accident modeling. *Accident Analysis and Prevention*, *30*(4), 425–433.

Ledesma, F. J., Navarro, M., & Perez-Rodriguez, J. V. (2005). Return to tourist destination. Is it reputation, after all? *Applied Economics*, *37*, 2055–2065.

Miori, V. M. (2006). A novel approach to the continuous flow truckload routing problem. In: K. Lawrence & R. Klimberg (Eds), *Applications of management science: In productivity, finance, and operations* (Vol. 12, pp. 145–155). The Netherlands: Amsterdam.

Poch, M., & Mannering, F. (1996). Negative binomial analysis of intersection-accident frequencies. *Journal of Transportation Engineering*, 105–113.

Powell, W. B. (1996). A stochastic formulation of the dynamic assignment problem, with an application to truckload motor carriers. *Transportation Science*, *30*(3), 195–219.

Puig, P. (2003). Characterizing additively closed discrete models by a property of their MLEs, with an application to generalized hermite distributions. *Journal of the American Statistical Association*, *98*, 687–692.

Puig, P., & Valero, J. (2006). Count data distributions: Some characterizations with applications. *Journal of the American Statistical Association*, *101*(473), 332–340.

Yang, J., Jaillet, P., & Mahmassani, H. S. (2000). Study of a real-time multi-vehicle truckload pickup-and-delivery problem. Research funded by NSF grant DMI-9713682.

APPENDIX A. STATA OUTPUT: DESCRIPTIVE STATISTICS

```
Descriptive Statistics  Sunday April 23 21:08:00 2006   Page 1

                                                                  ___  ____  ____  ____  ____(tm)
                                                                 /__    /   ____/   /   ____/
                                                                ___/   /   /___/   /   /___/
                                                                  Statistics/Data Analysis

                                                                     User: Virginia M. Miori
                                                                  Project: Thesis

                                  tm
  /     /       /   /       /
 /   /   /   /   /   /   /    8.2   Copyright 1984-2005
  Statistics/Data Analysis           StataCorp
                                     4905 Lakeway Drive
     Special Edition                  College Station, Texas 77845 USA
                                     800-STATA-PC            http://www.stata.com
                                     979-696-4600            stata@stata.com
                                     979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  81980538203
         Licensed to:  Dorothy Brown
                       University of Pennsylvania

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

. insheet using c:\stata8\TLDataIn.csv
(20 vars, 4118 obs)

. gen TLStops= sumoftlstops-1

. summarize TLStops

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
     TLStops |      4118    .8244293    1.517293         0         14
```

APPENDIX B. STATA OUTPUT: REGRESSION RESULTS

```
Count Data Results 4/28/06  Thursday April 27 20:30:03 2006   Page 1

                                                   Statistics/Data Analysis

                                                     User: Virginia M. Miori
                                                   Project: Thesis

                                         8.2   Copyright 1984-2005
   Statistics/Data Analysis                     StataCorp
                                               4905 Lakeway Drive
      Special Edition                           College Station, Texas 77845 USA
                                               800-STATA-PC            http://www.stata.com
                                               979-696-4600            stata@stata.com
                                               979-696-4601 (fax)

Single-user Stata for Windows perpetual license:
       Serial number:  81980538203
         Licensed to:  Dorothy Brown
                       University of Pennsylvania

Notes:
      1.  (/m# option or -set memory-) 10.00 MB allocated to data
      2.  (/v# option or -set maxvar-) 5000 maximum variables

1 . insheet using c:\stata8\tldatain.csv
  (20 vars, 4116 obs)

2 . gen tlstops= sumoftlstops-1

3 . gen distsq= avgofmiles^2

4 . gen distcb= avgofmiles^3

5 . poisson  tlstops quarter2 quarter3 quarter4 window2 window3 origtonsout originindex avgofmiles des
  > ttonsin destindex distsq distcb

  Iteration 0:   log likelihood =   -5440.4648
  Iteration 1:   log likelihood =   -5439.6779
  Iteration 2:   log likelihood =   -5439.6763
  Iteration 3:   log likelihood =   -5439.6763

  Poisson regression                                Number of obs   =       4116
                                                       LR chi2(12)     =    1379.72
                                                       Prob > chi2     =     0.0000
  Log likelihood =  -5439.6763                         Pseudo R2       =     0.1125
```

tlstops	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
quarter2	-.1885313	.0526266	-3.58	0.000	-.2916776	-.085385
quarter3	-.1042429	.0490842	-2.12	0.034	-.2004462	-.0080396
quarter4	.0509443	.0478812	1.06	0.287	-.0429012	.1447898
window2	-.0393108	.0403377	-0.97	0.330	-.1183713	.0397496
window3	.0027213	.0432606	0.06	0.950	-.0820679	.0875105
origtonsout	3.24e-06	2.56e-07	12.67	0.000	2.74e-06	3.74e-06
originindex	-1.627377	.212564	-7.66	0.000	-2.043995	-1.210759
avgofmiles	-.0002644	.0003204	-0.83	0.409	-.0008924	.0003636
desttonsin	-3.00e-06	2.59e-07	-11.61	0.000	-3.51e-06	-2.50e-06
destindex	1.536877	.1261726	12.18	0.000	1.289584	1.784171
distsq	-1.71e-06	3.73e-07	-4.58	0.000	-2.44e-06	-9.78e-07
distcb	6.78e-10	1.10e-10	6.19	0.000	4.63e-10	8.93e-10
_cons	.1621269	.0787736	2.06	0.040	.0077335	.3165203

```
Count Data Results 4/28/06  Thursday April 27 20:30:15 2006   Page 2

6 . nbreg  tlstops quarter2 quarter3 quarter4 window2 window3 origtonsout originindex avgofmiles des
  > ttonsin destindex distsq distcb

Fitting Poisson model:

Iteration 0:   log likelihood =   -5440.4648
Iteration 1:   log likelihood =   -5439.6779
Iteration 2:   log likelihood =   -5439.6763
Iteration 3:   log likelihood =   -5439.6763

Fitting constant-only model:

Iteration 0:   log likelihood =   -5173.1576
Iteration 1:   log likelihood =   -5093.4215
Iteration 2:   log likelihood =   -5042.4705
Iteration 3:   log likelihood =   -5042.3588
Iteration 4:   log likelihood =   -5042.3588

Fitting full model:

Iteration 0:   log likelihood =   -4844.9068
Iteration 1:   log likelihood =   -4790.0312
Iteration 2:   log likelihood =   -4780.3201
Iteration 3:   log likelihood =   -4780.2732
Iteration 4:   log likelihood =   -4780.2732

Negative binomial regression                      Number of obs   =          4116
                                                    LR chi2(11)     =        524.17
                                                    Prob > chi2     =        0.0000
Log likelihood =  -4780.2732                        Pseudo R2       =        0.0520

------------------------------------------------------------------------------
     tlstops |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    quarter2 |   -.246382   .0818873    -3.01   0.003    -.4068782   -.0858859
    quarter3 |  -.1622518   .0776189    -2.09   0.037     -.314382   -.0101216
    quarter4 |  -.0076083   .0768814    -0.10   0.921    -.1582931    .1430764
     window2 |  -.0000437   .0638284    -0.00   0.999     -.125145    .1250576
     window3 |   .0184506   .0679844     0.27   0.786    -.1147964    .1516976
  origtonsout |   3.82e-06   4.52e-07     8.45   0.000     2.93e-06    4.71e-06
 originindex |  -1.324451   .3522491    -3.76   0.000    -2.014846   -.6340552
   avgofmiles |  -.0006165   .0005269    -1.17   0.242    -.0016492    .0004161
   desttonsin |  -2.67e-06   3.59e-07    -7.43   0.000    -3.37e-06   -1.96e-06
    destindex |   1.621209   .1894752     8.56   0.000     1.249844    1.992573
       distsq |  -1.16e-06   6.09e-07    -1.91   0.056    -2.36e-06    3.15e-08
       distcb |   5.18e-10   1.88e-10     2.75   0.006     1.49e-10    8.86e-10
        _cons |   .1677752   .1320141     1.27   0.204    -.0909676     .426518
-------------+----------------------------------------------------------------
      /lnalpha |   .4123711    .055842                      .3029227    .5218195
-------------+----------------------------------------------------------------
         alpha |   1.510395   .0843435                      1.35381    1.685091
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:    chibar2(01) =  1318.81 Prob>=chibar2 =  0.000

7 . nbreg  tlstops quarter2 quarter3 origtonsout originindex avgofmiles desttonsin destindex distsq
  > distcb

Fitting Poisson model:

Iteration 0:   log likelihood =   -5441.6742
Iteration 1:   log likelihood =    -5440.892
Iteration 2:   log likelihood =   -5440.8904
Iteration 3:   log likelihood =   -5440.8904

Fitting constant-only model:
```

```
Count Data Results 4/28/06  Thursday April 27 20:30:16 2006   Page 3

Iteration 0:   log likelihood =   -5173.1576
Iteration 1:   log likelihood =   -5093.4215
Iteration 2:   log likelihood =   -5042.4705
Iteration 3:   log likelihood =   -5042.3588
Iteration 4:   log likelihood =   -5042.3588

Fitting full model:

Iteration 0:   log likelihood =   -4844.1512
Iteration 1:   log likelihood =   -4789.6441
Iteration 2:   log likelihood =   -4780.3667
Iteration 3:   log likelihood =   -4780.3231
Iteration 4:   log likelihood =   -4780.3231

Negative binomial regression                      Number of obs   =          4116
                                                  LR chi2(8)      =        524.07
                                                  Prob > chi2     =        0.0000
Log likelihood =  -4780.3231                      Pseudo R2       =        0.0520

------------------------------------------------------------------------------
     tlstops |      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
    quarter2 |  -.2421635   .0702875    -3.45   0.001    -.3799244   -.1044025
    quarter3 |  -.1579375   .0652447    -2.42   0.015    -.2858148   -.0300602
  origtonsout |   3.82e-06   4.51e-07     8.47   0.000     2.94e-06    4.71e-06
  originindex |  -1.323618   .3517617    -3.76   0.000    -2.013058   -.634178
    avgofmiles |  -.0006154   .0005267    -1.17   0.243    -.0016479    .000417
   desttonsin |  -2.67e-06   3.59e-07    -7.44   0.000    -3.37e-06   -1.97e-06
    destindex |   1.620556   .1894467     8.55   0.000     1.249247    1.991864
        distsq |  -1.16e-06   6.09e-07    -1.90   0.057    -2.35e-06    3.46e-08
        distcb |   5.16e-10   1.88e-10     2.74   0.006     1.47e-10    8.85e-10
         _cons |   .1676728   .1197594     1.40   0.161    -.0670514     .402397
-------------+----------------------------------------------------------------
      /lnalpha |   .4124127   .0558234                      .3030008    .5218246
-------------+----------------------------------------------------------------
         alpha |   1.510458   .0843189                      1.353916    1.685099
------------------------------------------------------------------------------
Likelihood-ratio test of alpha=0:   chibar2(01) =  1321.13 Prob>=chibar2 =  0.000

 8 . predict count
   (option n assumed; predicted number of events)

 9 . outfile shipdate shiploc destloc tlstops count using "C:\Stata8\TLDemV2.csv", noquote replace co
   > mma

10 .
```

TWO-ATTRIBUTE WARRANTY POLICIES UNDER CONSUMER PREFERENCES OF USAGE AND CLAIMS EXECUTION

Amitava Mitra and Jayprakash G. Patankar

ABSTRACT

Warranty policies for certain products, such as automobiles, often involve consideration of two attributes, for example, time and usage. Since consumers are not necessarily homogeneous in their use of the product, such policies provide protection to users of various categories. In this chapter, product usage at a certain time is linked to the product age through a variable defined as usage rate. This variable, usage rate, is assumed to be a random variable with a specified probability distribution, which permits modeling of a variety of customer categories. Another feature of the chapter is to model the propensity to execute the warranty, in the event of a failure within specified parameter values (say time or usage). In a competitive market, alternative product/warranty offerings may reduce the chances of exercising the warranty. This chapter investigates the impact of warranty policy parameters with the goal of maximizing market share, subject to certain constraints associated with expected warranty costs per unit not exceeding a desirable level.

Advances in Business and Management Forecasting, Volume 6, 217–235

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006013

INTRODUCTION

A majority of consumer products provide some sort of assurance to the consumer regarding the quality of the product sold. This assurance, in the form of a warranty, is offered at the time of sale. The Magnuson–Moss Warranty Act of 1975 (US Federal Trade Commission Improvement Act, 1975) also mandates that manufacturers must offer a warranty for all consumer products sold for more than 15 dollars. The warranty statement assures consumers that the product will perform its function to their satisfaction up to a given amount of time (i.e., warranty period) from the date of purchase. Manufacturers offer many different types of warranties to promote their products. Thus, warranties have become a significant promotional tool for manufacturers. Warranties also limit the manufacturers' liability in the case of product failure beyond warranty period.

Taxonomy of the different types of warranty policies may be found in the work of Blischke and Murthy (1994). Considering warranty policies that do not involve product development after sale, policies exist for a single item or for a group of items. With our focus on single items, policies may be subdivided into the two categories of nonrenewing and renewing. In a renewing policy, if an item fails within the warranty time, it is replaced by a new item with a new warranty. In effect, warranty beings anew with each replacement. On the other hand, for a nonrenewing policy, replacement of a failed item does not alter the original warranty. Within each of these two categories, policies may be subcategorized as simple or combination. Examples of a simple policy are those that incorporate replacement or repair of the product, either free or on a pro rata basis. The proportion of the warranty time that the product was operational is typically used as a basis for determining the cost to the customer for a pro rata warranty. Given limited resources, management has to budget for warranty repair costs and thereby determine appropriate values of the warranty parameters of, say, time and usage.

Although manufacturers use warranties as a competitive strategy to boost their market share, profitability, and image, they are by no means cheap. Warranties cost manufacturers a substantial amount of money. The cost of a warranty program must be estimated precisely and its effect on the firm's profitability must be studied. Manufacturers plan for warranty costs through the creation of a fund for warranty reserves. These funds are set aside at the beginning of the sales period to meet product replacement or repair obligations that occur while the product is under warranty. An estimate of the expected warranty costs is thus essential for management to

plan for warranty reserves. For the warranty policy considered, we assume that the product will be repaired if failure occurs within a specified time and the usage is less than a specified amount. Such a two-dimensional policy is found for products such as automobiles where the warranty coverage is provided for a time period; say five years, and a usage limit of, say, 50,000 miles. In this chapter, we assume minimal repair, that is, the failure rate of the product on repair remains the same as just before failure. Further, the repair time is assumed to be negligible.

LITERATURE REVIEW

Estimation of warranty costs has been studied extensively in the literature. Menke (1969) estimated expected warranty costs for a single sale for a linear pro rata and lump-sum rebate plan for nonrenewable policies. An exponential failure distribution was assumed. A drawback of his study was that he ignored the time value of money. Since warranty costs are incurred in the future, it is important that they be discounted using a proper discounting rate. Amato and Anderson (1976) extended Menke's model to allow for discounting and price adjustments. Discounting may cause the product price to fall in real terms and be more competitive in the marketplace. Whereas these models estimate costs based on a selected warranty period, the formulation by Lowerre (1968) was motivated by the manufacturer's concern. His model determined the warranty time assuming that a selected percentage of revenue is used to meet warranty claims, thus providing the manufacturer with some guidelines for selecting the warranty period. Heschel (1971) developed his model with the consumer in mind. He found the expected repair cost to the consumer over the life of the product. His model assumed a full rebate for an initial period followed by a pro rata rebate.

Blischke and Scheuer (1975) considered the costs of two types of warranty policies, namely, the free-replacement and pro rata policy under different time-to-failure distributions. In a separate paper, Blischke and Scheuer (1981) applied renewal theory to estimate warranty costs for the two types of renewable warranty policies. They assumed that the buyer purchases an identical replacement when an item in service fails at the end of the warranty period. Patankar and Worm (1981) developed prediction intervals for warranty reserves and cash flows associated with linear pro rata and lump sum rebate plans. They developed confidence intervals on total warranty reserves and cash flows. Upper bounds on cash flows and warranty reserves

are determined to analyze the uncertainty or risk involved. Using a Markovian approach, Balachandran, Maschmeyer, and Livingstone (1981) estimated the repair and replacement cost of a warranty program. They assumed an exponential failure distribution since its failure rate is constant.

Mamer (1982) estimated short-run total costs and long-run average costs of products under warranty. He studied pro rata and free-replacement warranty policies under different failure distributions (such as new-better-than-used) and showed that the expected warranty costs depends on the mean of the product lifetime distribution and its failure rate. Mamer considered pro rata and free-replacement warranties. He assumed that a customer purchases a new product instantaneously during the product life cycle. This provides a good assumption for using renewal theory. Cost of the replacement depends on the type of rebate offered. Mamer (1987) later expanded his previous research with present value analysis and analyzed the trade-off between warranty and quality control. He considered three warranty policies: ordinary replacement (OR), free replacement (FR), and pro rata replacement (PR). In OR, an item is replaced and warranty is offered only up to the original warranty period. In FR, a new warranty is offered when an item is replaced. The new warranty is the same as the original warranty. In PR, the consumer is charged based on time of failure, and a new warranty, the same as the original is offered.

Blacer and Sahin (1986) estimated warranty costs under free-replacement and pro rata warranty policies using renewal theory. They used the concept of a renewal process to estimate warranty costs over the product life cycle for pro rata and free-replacement warranties and computed first and second cost moments using gamma, exponential, and mixed exponential failure distributions. Frees and Nam (1988) found expressions for expected warranty costs in terms of distribution functions. They considered the free-replacement and pro rata policies to estimate warranty costs, where warranties are renewed under certain conditions. They concluded that it is very difficult to estimate renewal functions, and, therefore, warranty costs mathematically and suggest a couple of approximations. They used new better than used (NBU) distribution and straight-line approximation (SLA) methods. They found that SLA gives a very good approximation to the estimations provided by Nguyen and Murthy (1984a, 1984b). The drawback of Nguygen and Murthy's model was that they assumed a monotonic failure rate for lifetime distributions.

Frees (1988) showed that estimating warranty costs is similar to estimating renewal functions in renewal theory. He also estimated the variability of warranty costs using parametric and nonparametric methods.

Tapiero and Posner (1988) presented an alternative approach to modeling of warranty reserves. Thomas (1989) found the expected warranty reserve cost per unit with discounting for failure distributions that included the uniform, gamma, and Weibull. Patankar and Mitra (1995) studied the effect of partial execution of warranty and its impact on warranty costs to model different factors that influence consumer behavior in exercising their warranty rights. Mitra and Patankar (1997) considered warranty programs that offer customers the option to renew warranty, after an initial warranty period, for a certain premium. The effect of such warranty programs on market share and warranty costs is investigated.

A good review of the various warranty policies is found in Blischke and Murthy (1992). Murthy and Blischke (1992a) provide a comprehensive framework of analyses in product warranty management and further conduct a detailed review of mathematical models (Murthy & Blischke, 1992b) in this research area. A thorough treatment of warranty cost models and analysis of specific types of warranty policies, along with operational and engineering aspects of product warranties, is found by Blischke and Murthy (1994). The vast literature in warranty analysis is quite disjoint. A gap exists between researchers from different disciplines. With the objective of bridging this gap, Blischke and Murthy (1996) provided a comprehensive treatise of consumer product warranties viewed from different disciplines. In addition to providing a history of warranty, the handbook presents topics such as warranty legislation and legal actions; statistical, mathematical, and engineering analysis; cost models; and the role of warranty in marketing, management, and society.

Murthy and Djamaludin (2002) provided a literature review of warranty policies for new products. As each new generation of product usually increases in complexity to satisfy consumer needs, customers are initially uncertain about its performance and may rely on warranties to influence their product choice. Additionally, servicing of warranty, whether to repair or replace the product by a new one, influences the expected cost to the manufacturer (Jack & Murthy, 2001). Wu, Lin, and Chou (2006) considered a model for manufacturers to determine optimal price and warranty length to maximize profit based on a chosen life cycle for a free renewal warranty policy. Huang, Liu, and Murthy (2007) developed a model to determine the parameters of product reliability, price, and warranty strategy that maximize integrated profit for repairable products sold under a FR repair warranty strategy.

The majority of past research has dealt with a single-attribute warranty policy, where the warranty parameter is typically the time since purchase of

the product. Singpurwalla (1987) developed an optimal warranty policy based on maximization of expected utilities involving both profit and costs. A bivariate probability model involving time and usage as warranty criteria was incorporated. One of the first studies among two-dimensional warranty policies using a one-dimensional approach is that by Moskowitz and Chun (1988). Product usage was assumed to be a linear function of the age of the product. Singpurwalla and Wilson (1993, 1998) model time to failure, conditional on total usage. By choosing a distribution for total usage, they derive a two-dimensional distribution for failure using both age and usage. Singpurwalla (1992) also considers modeling survival in a dynamic environment with the usage rate changing dynamically. Moskowitz and Chun (1994) used a Poisson regression model to determine warranty costs for two-dimensional warranty policies. They assumed that the total number of failures is Poisson distributed whose parameter can be expressed as a regression function of age and usage of a product. Murthy, Iskander, and Wilson (1995) used several types of bivariate probability distributions in modeling product failures as a random point process on the two-dimensional plane and considered free-replacement policies. Eliashberg, Singpurwalla, and Wilson (1997) considered the problem of assessing the size of a reserve needed by the manufacturer to meet future warranty claims in the context of a two-dimensional warranty. They developed a class of reliability models that index failure by two scales, such as time and usage. Usage is modeled as a covariate of time. Gertsbakh and Kordonsky (1998) reduced usage and time to a single scale, using a linear relationship. Ahn, Chae, and Clark (1998) used a similar concept using a logarithmic transformation.

Chun and Tang (1999) found warranty costs for a two-attribute warranty model by considering age and usage of the product as warranty parameters. They provided warranty cost estimation for four different warranty policies; rectangular, L-shaped, triangular, and iso-cost, and performed sensitivity analysis on discount rate, usage rate, and warranty terms to determine their effects on warranty costs. Kim and Rao (2000) considered a two-attribute warranty model for nonrepairable products using a bivariate exponential distribution to explain item failures. Analytical expressions for warranty costs are derived using Downtone's bivariate distribution. They demonstrate the effect of correlation between usage and time on warranty costs. A two-dimensional renewal process is used to estimate warranty costs. Hsiung and Shen (2001) considered the effect of warranty costs on optimization of the economic manufacturing quality (EMQ). As a process deteriorates over time, it produces defective items that incur reworking costs (before sale) or

warranty repair costs (after sale). The objective of their paper is to determine the lot size that will minimize total cost per unit of time. The total cost per unit of time includes set up cost, holding cost, inspection cost, reworked cost, and warranty costs. Sensitivity analysis is performed on various costs to determine an optimum production lot size. Yeh and Lo (2001) explored the effect of preventive maintenance actions on expected warranty costs. A model is developed to minimize such costs. Providing a regular preventive maintenance within the warranty period increases maintenance cost to the seller, but the expected warranty cost is significantly reduced. An algorithm is developed that determines an optimal maintenance policy. Lam and Lam (2001) developed a model to estimate expected warranty costs for a warranty that includes a free repair period and an extended warranty period. Consumers have an option to renew warranty after the free repair period ends. The choice of consumers has a significant effect on the expected warranty costs and determination of optimal warranty policy.

The model developed in this chapter is unique from those considered in the past in several aspects. In the context of a two-attribute policy (say time and usage) where a failed item is repaired or replaced, to model a variety of consumers, usage rate is considered to be a random variable with a probability distribution. Another unique contribution is the modeling of the customer's propensity to exercise warranty, in the event of a product failure within the stipulated parameters of the warranty policy. This is incorporated through the development of a warranty execution function, which represents a more realistic model. Since all consumers may not necessarily exercise warranty claims, the impact of this warranty execution function will lead to a downward adjustment in the expected warranty costs. This, in turn, will reflect a smaller accrued liability on the financial statements of the firm.

The model considered in this chapter is also unique from the others in its ability to deal with current markets (that also includes gray markets) and enterprise warranty programs. Gray-market products are those that are offered for sale by unauthorized third parties. Although gray market products are not illegal, since they are sold outside of the normal distribution channels of the original equipment manufacturer (OEM), a manufacturer may not honor the warranty. On the one hand, although consumers may buy the product at a lower purchase price in the grey market, the assurance of receiving warranty-related service is slim. Hence, in case of product failure within the warranty period, customer satisfaction with the OEM will be hurt, impacting future sales and market share of the OEM.

By incorporating a "warranty execution function" in the formulation (described later by Eqs. (5)–(8)), the chapter has the ability to model "fewer" executions of warranty, in case the product is obtained through the gray market. In particular, two parameters, the time up to which full warranty execution takes place (t_c), and the warranty attrition rate (δ), may be adequately selected to model product purchased from gray markets. If the OEM has an idea of the proportion of sales that is from gray markets, it will assist in a better forecast of expected warranty costs. When the OEM definitely does not honor warranty for gray market products, it will force the parameter t_c to be zero. Alternatively, if the OEM provides limited warranty support and service, this may lead to a choice of the parameter t_c to be some small value, greater than zero.

The choice of the warranty execution attrition rate parameter, δ, is also influenced if the product is from gray markets. When the consumer is suspect about the warranty being honored, it leads to selection of δ that permits rapid attrition. Thus, small values of δ could be used to model this situation, compared to the case where the product is purchased through authorized distributors or retailers.

Another feature of the existing chapter is the ability to model enterprise warranty programs. Such programs are generally flexible and more generous and provide service and support that extends the original warranty coverage. Associated with the enterprise warranty program is the emergence of service contract providers that underwrite these policies. These providers manage the claims processing through a warranty management system that is timely and responsive to the customer. Significant reduction in the time to process warranties may also take place, leading to improved customer satisfaction. The warranty execution function, incorporated in this model, can incorporate the impact of enterprise warranty programs on warranty costs. First, the time up to which full warranty execution takes place (t_c), will normally increase in such a program. Second, due to improved customer satisfaction associated with such enterprise programs, the warranty execution attrition rate parameter (δ) will be influenced such that there is a lower attrition. Hence, large values of the parameter δ could be appropriately selected to model such situations. Lastly, the probability distribution of the parameter, t_c, could be affected by such programs. The mean value of t_c tends to be larger. The variability of the distribution of t_c could be smaller, compared to a regular warranty program, due to consistency in customer satisfaction. This can be modeled through adequate selection of the parameters α and β (as represented by Eq. (6)).

Research Objectives

In this chapter, we consider a two-dimensional warranty policy where the warranty parameters, for example, could be time and usage at the point of product failure. A warranty policy in this context, such as those offered for automobiles, could be stated as follows: product will be replaced or repaired free of charge up to a time (W) or up to a usage (U), whichever occurs first from the time of the initial purchase. Warranty is not renewed on product failure. For example, automobile manufacturers may offer a 36 months or 36,000 miles warranty, whichever occurs first. For customers with high usage rates, the 36,000 miles may occur before 36 months. On the contrary, for those with limited usage, the warranty time period of 36 months may occur first. Fig. 1 shows a two-dimensional warranty region.

We assume that the usage is related to time as a linear function through the usage rate. To model a variety of consumers, usage rate is assumed to be a random variable with a specified probability distribution. This chapter develops a model based on minimal – repair or replacement of failed items. Another feature of this chapter is to incorporate the situation that warranty may not be executed all the time. For example, a customer may develop a dissatisfaction for the product and prefer to switch brands, rather than to exercise the warranty. Instances such as lost warranties, customer relocation, and inconvenience in getting the product repaired may also dissuade a customer from exercising warranty.

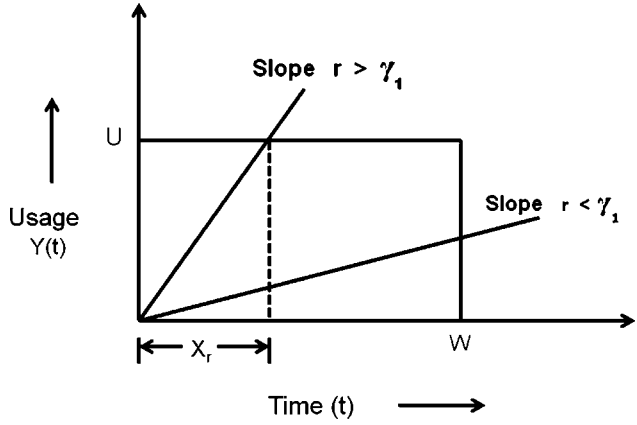

Fig. 1. Two-Dimensional Warranty Region.

In this chapter, we develop a model from the manufacturer's perspective. Using consumer conditions of usage and preferences on execution of warranty claims in the event of product failure within the specified bounds of warranty parameters, expected warranty costs are determined. The manufacturer typically has an idea of the maximum bound, of expected warranty costs to sales, that should not be exceeded. Using this as a constraint, in addition to upper and lower bounds on the price, warranty time, and usage, optimal parameter values are determined based on maximizing market share.

MODEL DEVELOPMENT

The following notation is used in the chapter:

W	Warranty period offered in warranty policy
U	Usage limit offered in warranty policy
R	Usage rate
t	Instant of time
$Y(t)$	Usage at time t
$\lambda(t\|r)$	Failure intensity function at time t given $R = r$
t_c	Time up to which full warranty execution takes place
$m(t\|t_c)$	Conditional warranty execution function at time t given t_c
$q(t_c)$	Probability density function of t_c
$N(W, U\|r)$	Number of failures under warranty given $R = r$
c	Unit product price
c_s	Unit cost of repair or replacement
Q	Market share for a given policy

Relationship Between Warranty Attributes

We assume that the two attributes, say time and usage, are related linearly through the usage rate, which is a random variable. Denoting $Y(t)$ to be the usage at time t and $X(t)$ the corresponding age, we have

$$Y(t) = RX(t) \tag{1}$$

where R is the usage rate. It is assumed that all items that fail within the prescribed warranty parameters are minimally repaired and the repair time is negligible. In this context, $X(t) = t$.

Distribution Function of Usage Rage

To model a variety of customers, R is assumed to be a random variable with probability density function given by $g(r)$. The following distribution functions of R are considered in this chapter:

(a) R has a uniform distribution over (a_1, b_1):
This models a situation where the usage rate is constant across all customers. The density function of R is given by

$$\begin{aligned} g(r) &= \frac{1}{b_1 - a_1}, \quad a_1 \le r \le b_1 \\ &= 0, \quad \text{otherwise} \end{aligned} \tag{2}$$

(b) R has a gamma distribution function:
This may be used for modeling a variety of usage rates among the population of consumers. The shape of the gamma distribution function is influenced by the selection of its parameters. When the parameter, p, is equal to 1, it reduces to the exponential distribution. The density function is given by

$$g(r) = \frac{e^{-r} r^{p-1}}{\Gamma(p)}, \quad 0 \le r < \infty, p > 0 \tag{3}$$

Failure Rate

Failures are assumed to occur according to a Poisson process where it is assumed that failed items are minimally repaired. If the repair time is small, it can be approximated as being zero. Since the failure rate is unaffected by minimal repair, failures over time occur according to a nonstationary Poisson process with intensity function $\lambda(t)$ equal to the failure rate.

Conditional on the usage rate $R = r$, let the failure intensity function at time t be given by

$$\lambda(t|r) = \theta_0 + \theta_1 r + (\theta_2 + \theta_3 r)t \tag{4}$$

(1) Stationary Poisson process:
Under this situation, the intensity function $\lambda(t|r)$ is a deterministic quantity as a function of t when $\theta_2 = \theta_3 = 0$. This applies to many electronic components that do not deteriorate with age and failures are due to pure chance. The failure rate in this case is constant.

(2) Nonstationary Poisson process:
This models the more general situation where the intensity function changes as a function of t. It is appropriate for products and components with moving parts where the failure rate may increase with time of usage. In this case θ_2 and θ_3 are not equal to zero.

Warranty Execution Function

A variety of reasons, as mentioned previously, may prevent the full execution of warranty. The form of the weight function, which describes warranty execution, could be influenced by factors such as the warranty time, usage limit, warranty attrition due to costs of executing the warranty, and the product class as to whether they are expensive.

Fig. 2 shows a conditional warranty execution function, $m(t|t_c)$. It is assumed that full execution will take place if $t \leq t_c$. Beyond t_c, the conditional execution function decreases exponentially with t. Eq. (5) represents $m(t|t_c)$:

$$\begin{aligned} m(t|t_c) &= 1, && 0 < t \leq t_c \\ &= \exp\left[-(t - t_c)/\delta\right], && t_c \leq t \leq W \\ &= 0, && t > W \end{aligned} \tag{5}$$

The parameter δ in Eq. (5) is a measure of the rate of warranty attrition due to some of the reasons discussed previously. As δ increases, the rate of warranty attrition decreases, leading to an increase in the value of the conditional warranty execution function.

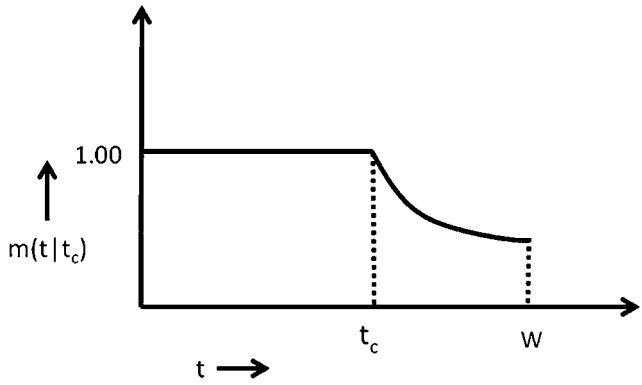

Fig. 2. Conditional Warranty Execution Function.

Note that if $t_c = W$, we have full warranty execution. To model heterogeneity in customer behavior in warranty execution, the distribution of t_c is assumed to be given by a gamma distribution

$$q(t_c) = \frac{\beta^{\alpha}}{\Gamma(\alpha)} \exp(-\beta t_c) t_c^{\alpha-1} \tag{6}$$

The choice of the parameter values of α and β, that determine the distribution of t_c will be influenced by the current market and enterprise environment. Note that for a gamma distribution the mean and variance are given by α/β and α/β^2, respectively. For products purchased in gray markets, the mean of t_c will be small, implying small values of α and large values of β. Alternatively, for enterprise warranty programs, the mean of t_c will be larger, relative to the traditional warranty program. This may be accomplished through selection of large values of α and small values of β. However, since the variability in the distribution of t_c will be smaller than that of a traditional warranty program, a judicious choice of α and β will have to be made. Although increasing the value of β will reduce the variance of t_c, it will also reduce the mean of t_c. Thus, based on market conditions, α and β may have to be jointly selected such that the existing environment on the mean and variance of t_c are modeled appropriately.

Now, $m(t)$, the warranty execution function is found by compounding $m(t|t_c)$ with $q(t_c)$ as follows:

$$m(t) = \int m(t|t_c) q(t_c) dt_c \tag{7}$$

Here, the exponential distribution, which is a special case of the gamma distribution (when $\alpha = 1$) is used for $q(t_c)$ to derive the results. Thus, we have

$$\begin{aligned} m(t) &= \int_{t_c=0}^{t} \beta \exp\left[-(t - t_c)/\delta - \beta t_c\right] dt_c + \int_{t_c=t}^{W} \beta \exp[-(\beta t_c)] dt_c \\ &= [\exp(-\beta t) - \exp(-\beta W)] + \frac{\beta \exp(-t/\delta)}{(\beta - 1/\delta)} [1 - \exp[-(\beta - 1/\delta)t]] \end{aligned} \tag{8}$$

Market Share

The market share function (Q) is formulated so as to be bounded between 0 and 1. It is developed to decrease exponentially with respect to product price (c), increase exponentially with warranty time (W) as well as usage

limit (U), and is given by

$$Q = Dc^{-a}(W + k)^b U^d \tag{9}$$

where a is the constant representing the price elasticity of the product, $a > 1$; k the constant of warranty time displacement allowing for the possibility of nonzero market share when warranty time is 0; b the constant representing the displaced warranty period elasticity of the product, $0 < b < 1$; d the constant representing the warranty usage limit elasticity of the product, $0 < d < 1$; and D the normalizing constant.

Assuming that the manufacturer has an idea of the maximum possible market share (D_1) that may be attained by the product, the normalizing constant (D), is given by

$$D = D_1/[c_1^{-a}(W_2 + k)^b U_2^d] \tag{10}$$

where c_1 is the lower bound on product price, W_2 the upper bound on warranty time, and U_2 the upper bound on usage limit.

Expected Warranty Costs

The warranty region is the rectangle shown in Fig. 1, where W is the warranty period and U the usage limit. Let $\gamma_1 = U/W$. Conditional on the usage rate $R = r$, if the usage rate $r \geq \gamma_1$, warranty ceases at time X_r is given by

$$X_r = U/r \tag{11}$$

Alternatively, if $r < \gamma_1$, warranty ceases at time W. The number of failures under warranty conditional on $R = r$ is given by

$$N(W, U|r) = \begin{cases} \int_{t=0}^{W} \lambda(t|r)m(t)dt, & \text{if } r < \gamma_1 \\ \int_{t=0}^{X_r} \lambda(t|r)m(t)dt, & \text{if } r \geq \gamma_1 \end{cases} \tag{12}$$

The expected number of failures is thus obtained from

$$\begin{aligned} E[N(W, U)] = &\int_{r=0}^{\gamma_1} \left[\int_{t=0}^{W} \lambda(t|r)m(t)dt \right] g(r)dr \\ &+ \int_{r=\gamma_1}^{\infty} \left[\int_{t=0}^{X_r} \lambda(t|r)m(t)dt \right] g(r)dr \end{aligned} \tag{13}$$

Expected warranty costs (EWC) per unit are, therefore, given by

$$EWC = c_s E[N(W, U)] \tag{14}$$

whereas the expected warranty costs per unit sales (ECU) are obtained from

$$\text{ECU} = (c_s/c)E[N(W, U)] \tag{15}$$

Mathematical Model

We first consider the constraints that must be satisfied for the decision variables of product price, warranty time, and warranty usage limit. A manufacturer having knowledge of the unit cost of production and a desirable profit margin can usually identify a minimum price, below which it would not be feasible to sell the product. Similarly, knowing the competition, it has a notion of the maximum price that the product should be priced at. Using a similar rationale, a manufacturer might be able to specify minimum and maximum bounds on the warranty time and usage limit to be offered with the product. So, the constraints on the policy parameters are

$$\begin{aligned} c_1 \le c \le c_2 \\ W_1 \le W \le W_2 \\ U_1 \le U \le U_2 \end{aligned} \tag{16}$$

where c_1 is the minimum product price, c_2 the maximum product price, W_1 the minimum warranty period, W_2 the maximum warranty period, U_1 the minimum usage limit, and U_2 the maximum usage limit.

The manufacturer has an objective of maximizing market share. However, the manufacturer cannot indefinitely increase the warranty time or usage limit to do so, since expected warranty costs increase with the values of these parameters. A manufacturer may typically be constrained on the expected warranty costs per unit sales. Hence, the model is formulated as follows:

$$\begin{aligned} &\max \text{Q} \\ &\text{s.t.} \\ &\text{ECU} \le \alpha_1 \end{aligned} \tag{17}$$

and Eq. (16) on the parameter constraints.

RESULTS

The application of the proposed model is demonstrated through some sample results using selected values of the model parameters. Owing to the

complexity of Eq. (13), which influences Eq. (15) and the associated constraint in the model, closed form solutions for the integrals are usually not obtainable. Numerical integration methods are used.

The failure rate intensity function, conditional on R, is selected to be a stationary Poisson process with parameters $\theta_2 = 0$, $\theta_3 = 0$, $\theta_0 = 0.005$, and $\theta_1 = 0.01$. The ratio of c_s/c is selected as 0.5. The distribution of the usage rate, R, is chosen to be uniform in the interval (1, 6). Bounds on the warranty policy parameters are as follows: unit product price between \$10,000 and \$60,000 ($c_1 = 1$, $c_2 = 6$); warranty period between 0.5 and 5 years ($W_1 = 0.5$, $W_2 = 5.0$); and usage limit between 10,000 and 60,000 miles ($U_1 = 1$, $U_2 = 6$). For the upper bound on expected warranty costs per unit sales, the parameter α_1 is chosen to be 0.1. The market share function is developed using the following parameter values: $a = 2$, $k = 1.0$, $b = 0.2$, $d = 0.1$, and $D_1 = 0.2$.

For the conditional warranty execution, the attrition parameter δ is selected to be 1, 3, and 5, respectively. Modeling of consumer behavior through the distribution of the parameter t_c, the time up to which full execution takes place, is achieved by a gamma distribution with parameters $\alpha = 1$ and $\beta = 2, 4$.

Table 1 shows some results on the optimal warranty policy parameters of price, warranty time, and usage for chosen values of the warranty execution function parameters. The corresponding market share is also depicted. For a given value of the parameter β (which influences the distribution of the time, t_c, up to which full execution takes place), as the parameter δ increases, the attrition to execute warranty decreases, implying a higher propensity to claim warranty in case of product failure within the bounds of the policy. It is observed that for small values of δ, the policy parameters (c, W, and U) are close to their lower, upper, and upper bounds, respectively. For $\beta = 2$ and $\delta = 1$, the optimal parameter values are $c = 1$, $W = 5$, and $U = 5$ for a market share of 19.6%. Note that c, W, and U are at their respective lower,

Table 1. Optimal Warranty Policy Parameters.

β	δ	c	W	U	Q
2	1	1	5	5	0.196
	3	1	5	2	0.179
	5	1	5	1.5	0.174
4	1	1	5	6	0.200
	3	1	5	2.1	0.180
	5	1	5	1.6	0.175

upper, and upper bounds, respectively. As δ increases, for a given β, the optimal value of U moves away from its upper bound. If no constraint had been placed on ECU, the market share could be improved by increasing the usage limit further. But, since the company has to manage its expected warranty costs, the constrained optimization problem leads to a more manageable and feasible solution. As the warranty execution parameter, δ, increases, since the chances of warranty execution increases, expected warranty costs will go up. Thus, management may not be able to simultaneously increase both the warranty period or the usage limit to their upper bounds, so as to contain the expected warranty costs per unit sales below the chosen bound. As observed from Table 1, with an increase in the value of δ, the usage limit (U) is found to decrease, which subsequently results in a smaller market share.

CONCLUSIONS

The chapter has considered a two-dimensional warranty policy with the objective of assisting the manufacturer in selecting parameter values of unit product price, warranty time, and usage limit. Since manufacturers operate within resource constraints, it usually has to plan for expected warranty costs. Such costs are influenced by the warranty period and usage limit. Thus, although it may be desirable to increase the values of these parameters, to increase market share relative to competitors, it is not always feasible to do so based on expected warranty costs per unit.

An important insight has been the inclusion of the propensity to claim a warranty on part of the consumer. Consumer behavior is influenced by several factors such as new product development, products offered by competitors, the ease or difficulty of filing a claim, and the cost associated with such filing, among others. Additionally, since consumers may vary in their usage of the product, the usage rage has been modeled using a random variable with a selected probability distribution.

Several avenues of extension to this research exist. One could explore the study of enterprise warranty programs where an expanded warranty policy is offered to the customer at the time of purchase or even thereafter. Such policies are purchased through the payment of a premium and could be administered by the manufacturer or typically a third party. The impact of these enterprise warranty programs is to increase market share. However, expected warranty costs over the entire range of the policy also increases. Thus, a net revenue per unit sales approach could be pursued to determine the optimal warranty policy parameters in this context.

REFERENCES

Ahn, C. W., Chae, K. C., & Clark, G. M. (1998). Estimating parameters of the power law process with two measures of failure rate. *Journal of Quality Technology*, *30*, 127–132.

Amato, N. H., & Anderson, E. E. (1976). Determination of warranty reserves: An extension. *Management Science*, *22*(12), 1391–1394.

Balachandran, K. R., Maschmeyer, R. A., & Livingstone, J. L. (1981). Product warranty period: A Markovian approach to estimation and analysis of repair and replacement costs. *Accounting Review*, *56*, 115–124.

Blacer, Y., & Sahin, I. (1986). Replacement costs under warranty: Cost moments and time variability. *Operations Research*, *34*, 554–559.

Blischke, W. R., & Murthy, D. N. P. (1992). Product warranty management – I: A taxonomy for warranty policies. *European Journal of Operations Research*, *62*, 127–148.

Blischke, W. R., & Murthy, D. N. P. (1994). *Warranty cost analysis*. New York: Marcel Dekker, Inc.

Blischke, W. R., & Murthy, D. N. P. (Eds). (1996). *Product warranty handbook*. New York: Marcel Dekker, Inc.

Blischke, W. R., & Scheuer, E. M. (1975). Calculation of the warranty cost policies as a function of estimated life distributions. *Naval Research Logistics Quarterly*, *22*(4), 681–696.

Blischke, W. R., & Scheuer, E. M. (1981). Applications of renewal theory in analysis of free-replacement warranty. *Naval Research Logistics Quarterly*, *28*, 193–205.

Chun, Y. H., & Tang, K. (1999). Cost analysis of two-attribute warranty policies based on the product usage rate. *IEEE Transactions on Engineering Management*, *46*(2), 201–209.

Eliashberg, J., Singpurwalla, N. D., & Wilson, S. P. (1997). Calculating the warranty reserve for time and usage indexed warranty. *Management Science*, *43*(7), 966–975.

Frees, E. W. (1988). Estimating the cost of a warranty. *Journal of Business and Economic Statistics*, *1*, 79–86.

Frees, E. W., & Nam, S. H. (1988). Approximating expected warranty cost. *Management Science*, *43*, 1441–1449.

Gertsbakh, I. B., & Kordonsky, K. B. (1998). Parallel time scales and two-dimensional manufacturer and individual customer warranties. *IIE Transactions*, *30*, 1181–1189.

Heschel, M. S. (1971). How much is a guarantee worth? *Industrial Engineering*, *3*(5), 14–15.

Hsiung, C., & Shen, S. H. (2001). The effects of the warranty cost on the imperfect EMQ model with general discrete shift distribution. *Production Planning and Control*, *12*(6), 621–628.

Huang, H. Z., Liu, Z. J., & Murthy, D. N. P. (2007). Optimal reliability, warranty and price for new products. *IIE Transactions*, *39*, 819–827.

Jack, N., & Murthy, D. N. P. (2001). Servicing strategies for items sold with warranty. *Journal of Operational Research*, *52*, 1284–1288.

Kim, H. G., & Rao, B. M. (2000). Expected warranty cost of two-attribute free replacement warranties based on a bivariate exponential distribution. *Computers and Industrial Engineering*, *38*, 425–434.

Lam, Y., & Lam, P. K. W. (2001). An extended warranty policy with options open to the consumers. *European Journal of Operational Research*, *131*, 514–529.

Lowerre, J. M. (1968). On warranties. *Journal of Industrial Engineering*, *19*(3), 359–360.

Mamer, J. W. (1982). Cost analysis of pro rata and free-replacement warranties. *Naval Research Logistics Quarterly*, *29*(2), 345–356.

Mamer, J. W. (1987). Discounted and per unit costs of product warranty. *Management Science*, *33*(7), 916–930.

Menke, W. W. (1969). Determination of warranty reserves. *Management Science*, *15*(10), 542–549.

Mitra, A., & Patankar, J. G. (1997). Market share and warranty costs for renewable warranty programs. *International Journal of Production Economics*, *50*, 155–168.

Moskowitz, H., & Chun, Y. H. (1988). *A Bayesian approach to the two-attribute warranty policy*. Paper No. 950. Krannert Graduate School of Management, Purdue University, West Lafayette, IN.

Moskowitz, H., & Chun, Y. H. (1994). A Poisson regression model for two-attribute warranty policies. *Naval Research Logistics*, *41*, 355–376.

Murthy, D. N. P., & Blischke, W. R. (1992a). Product warranty management – II: An integrated framework for study. *European Journal of Operations Research*, *62*, 261–281.

Murthy, D. N. P., & Blischke, W. R. (1992b). Product warranty management – III: A review of mathematical models. *European Journal of Operations Research*, *62*, 1–34.

Murthy, D. N. P., & Djamaludin, I. (2002). New product warranty: A literature review. *International Journal of Production Economics*, *79*, 231–260.

Murthy, D. N. P., Iskander, B. P., & Wilson, R. J. (1995). Two dimensional failure free warranty policies: Two dimensional point process models. *Operations Research*, *43*, 356–366.

Nguyen, D. G., & Murthy, D. N. P. (1984a). A general model for estimating warranty costs for repairable products. *IIE Transactions*, *16*(4), 379–386.

Nguyen, D. G., & Murthy, D. N. P. (1984b). Cost analysis of warranty policies. *Naval Research Logistics Quarterly*, *31*, 525–541.

Patankar, J. G., & Mitra, A. (1995). Effects of warranty execution on warranty reserve costs. *Management Science*, *41*, 395–400.

Patankar, J. G., & Worm, G. H. (1981). Prediction intervals for warranty reserves and cash flow. *Management Science*, *27*(2), 237–241.

Singpurwalla, N. D. (1987). *A strategy for setting optimal warranties*. Report TR-87/4. Institute for Reliability and Risk Analysis, School of Engineering and Applied Science, George Washington University, Washington, DC.

Singpurwalla, N. D. (1992). Survival under multiple time scales in dynamic environments. In: J. P. Klein & P. K. Goel (Eds), *Survival analysis: State of the art* (pp. 345–354).

Singpurwalla, N. D., & Wilson, S. P. (1993). The warranty problem: Its statistical and game theoretic aspects. *SIAM Review*, *35*, 17–42.

Singpurwalla, N. D., & Wilson, S. P. (1998). Failure models indexed by two scales. *Advances in Applied Probability*, *30*, 1058–1072.

Tapiero, C. S., & Posner, M. J. (1988). Warranty reserving. *Naval Research Logistics Quarterly*, *35*, 473–479.

Thomas, M. U. (1989). A prediction model for manufacturer warranty reserves. *Management Science*, *35*, 1515–1519.

US Federal Trade Commission Improvement Act. (1975). 88 *Stat 2183*, 101–112.

Wu, C. C., Lin, P. C., & Chou, C. Y. (2006). Determination of price and warranty length for a normal lifetime distributed product. *International Journal of Production Economics*, *102*, 95–107.

Yeh, R. H., & Lo, H. C. (2001). Optimal preventive maintenance warranty policy for repairable products. *European Journal of Operational Research*, *134*, 59–69.

This page intentionally left blank

A DUAL TRANSPORTATION PROBLEM ANALYSIS FOR FACILITY EXPANSION/ CONTRACTION DECISIONS: A TUTORIAL

N. K. Kwak and Chang Won Lee

ABSTRACT

A dual transportation analysis is considered as a strategic matter for plant facility expansion/contraction decision making in manufacturing operations. The primal-dual problem is presented in a generalized mathematical form. A practical technique of generating the dual solution is illustrated with a plant facility expansion/contraction example as a tutorial. Demand forecasting is performed based on the time series data with seasonal variation adjustments. The dual solution helps facilitate operations decision making by providing useful information.

1. INTRODUCTION

A plethora of information on plant facility location analysis has appeared in the existing literature (Klamroth, 2002; Sule, 1988). The location selection

Advances in Business and Management Forecasting, Volume 6, 237–248

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006014

methods can be categorized as following: (1) factor listing/scoring method, (2) analytical hierarchy process (AHP), (3) mathematical programming methods, (4) simulation methods, and (5) heuristic algorithms.

In the factor listing/scoring method, decision makers identify and select the factors favorable for business expansion and rank them (Keeney, 1994). The AHP, developed by Saaty (1980), is a method of comparing and ranking decision alternatives and select the best one in location and strategic acquisition decision making (Bowen, 1995; Tavana & Banerjee, 1995; Klorpela & Truominen, 1996; Badri, 1999). Mathematical programming refers to a group of mathematical techniques used for optimal solutions subject to a set of decision constraints in facility location analysis. It includes linear programming, integer programming, dynamic programming, multicriteria decision-making methods, and data envelopment analysis (DEA), (Kwak, 1973; Erlenkotten, 1978; Kwak & Schniederjans, 1985; Sinha & Sastry, 1987; Current, Min, & Schilling, 1990; Revelle & Laporte, 1996; Korhonen & Syrjanen, 2004; Drezner & Hamacher, 2004; Campbell, Ernst, & Krishnamoorthy, 2005; Farahani & Asgari, 2007; Ho, Lee, & Ho, 2008). Simulation methods refer to the collection of methodologies used in the Monte Carlo simulation process. They are generally computer-based approaches to decision making, which replicate behavior of an operations system (Mehrez, Sinuany-Stern, Arad-Geva, & Binyamin, 1996; Quarterman, 1998; Schniederjans, 1999). Heuristic algorithms employ the use of some intuitive rules or guidelines to generate new strategies, which yield improved solutions. Some of the heuristic algorithms that can be used for facility expansion analysis are Tabu search algorithm (Glover, 1990), and genetic algorithm (Jaramillio, Bhadury, & Batta, 2002).

In this study, the dual transportation-problem approach is considered for facility expansion/contraction decision making and demand forecasting with an illustrative case example.

2. A DUALITY ANALYSIS OF THE TRANSPORTATION PROBLEM

The transportation problem, as a special case of linear programming, has been analyzed in various forms in operations research/management science literature. Most existing analyses deal with only one aspect of the optimization procedure, known as the primal problem. In addition to the primal aspect, every transportation problem has another related aspect, known as the dual problem. The dual aspect of the transportation problem

reveals some implicit business and mathematical relations. Because of the unusual mathematical formulation, the dual aspect of the transportation problem has not been given adequate attention to many students in operations research/management science. This section presents the primal-dual aspect of the transportation problem with an illustrative example and examines the implicit relations contained in the dual problem.

2.1. Dual Problem Formulation

The transportation problem can be stated in the general form as

$$\min \quad Z = \sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} \tag{1}$$

s.t.

$$\begin{aligned} &\sum_{j=1}^{n} x_{ij} = a_i \\ &\sum_{i=1}^{m} x_{ij} = b_j \\ &\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j \end{aligned} \tag{2}$$

$$x_{ij} \geq 0 \quad (i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n)$$

where c_{ij} is the unit transportation cost from the ith plant to the jth destination, x_{ij} the amount of product shipped from the ith plant to the jth destination, a_i the amount of supply at the ith plant, and b_j the amount demanded at the jth destination

In this general form, the problem is assumed to be balanced. In the dual formulation, the equality side constraints are converted into inequalities as

$$\sum_{j=1}^{n} x_{ij} \leq a_i \quad \text{or} \quad \sum_{j=1}^{n} -x_{ij} \geq -a_i$$

and

$$\sum_{i=1}^{m} x_{ij} \geq b_j$$

These changes do not violate the supply and demand conditions, because each of the i plants cannot supply a greater amount than it produces and each of the j destinations receives at least the amount demanded.

The dual problem is formulated as

$$\max \quad Z' = \sum_{j=1}^{n} b_j z_j - \sum_{i=1}^{m} a_i y_i \tag{3}$$

$$\begin{aligned} &\text{s.t.} \\ &z_j - y_i \leq c_{ij} \end{aligned} \tag{4}$$

$$y_i, z_j \geq 0 \quad (i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n)$$

where y_i and z_j are the dual decision variables representing the supply and demand requirement restrictions, respectively.

Since $Z = Z'$ in the optimal solution, Eq. (5) can be obtained from Eqs. (1) and (3)

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} = \sum_{j=1}^{n} b_j z_j - \sum_{i=1}^{m} a_i y_i \tag{5}$$

Substitution of a_i and b_j values in Eq. (2) into Eq. (5) yields

$$\sum_{i=1}^{m} \sum_{j=1}^{n} c_{ij} x_{ij} + \sum_{i=1}^{m} (y_i \sum_{j=1}^{n} x_{ij}) - \sum_{j=1}^{n} (z_j \sum_{i=1}^{m} x_{ij}) = 0$$

or

$$\sum_{i=1}^{m} \sum_{j=1}^{n} x_{ij} (c_{ij} + y_i - z_j) = 0 \tag{6}$$

Since $x_{ij} \geq 0$ and $(c_{ij} + y_i - z_j) \geq 0$, the following relations can be found from Eq. (6):

$$\begin{aligned} z_j &= c_{ij} + y_i \quad \text{if} \quad x_{ij} > 0 \\ z_j &\leq c_{ij} + y_i \quad \text{if} \quad x_{ij} = 0 \end{aligned} \tag{7}$$

Here, the dual variables y_i and z_j are interpreted as the value of the product, free on board (fob) at the ith plant and its delivered value at the jth

destination, respectively. Thus, the delivered values at the destinations are equal to the values at the origins plus the unit transportation costs.

3. AN ILLUSTRATIVE CASE EXAMPLE

A logistics company in this study is a leading independent provider of factory-to-dealer transportation solutions in Korea. The company deals with various products and distributors, such as agricultural and related equipments. The company's mission statement is to give customers competitive advantages through logistics expertise, adequate market forecasting, and local knowledge. The company has four plants: Busan (A), Inchon (B), Kunsan (C), and Mokpo (D). There are five dealer–distributors in the different markets. The company distributes farm and related equipments by means of inland and sea transportation modes.

The company management is concerned about recent turbulence in the global economy, which affects the domestic market performance, especially due to the seasonally varied demand for the products. Therefore, the general managerial concern is to find the best practice in transportation services to the dealer–distributors while minimizing the total transportation cost, as well as fulfilling their needs. Data templates were derived from the company record with some modifications to ensure the corporate security in the competitive

Table 1. Transportation Table.

Dstn.* / Plant	I	II	III	IV	V	Supply
A	13 x_{11}	18 x_{12}	12 x_{13}	15 x_{14}	13 x_{15}	40
B	15 x_{21}	15 x_{22}	10 x_{23}	17 x_{24}	10 x_{25}	50
C	12 x_{31}	12 x_{32}	10 x_{33}	16 x_{34}	23 x_{35}	60
D	19 x_{41}	20 x_{42}	17 x_{43}	25 x_{44}	15 x_{45}	50
Demand	40	50	30	30	50	200

*Destination (Dstn) = Dealer-Distributor

market, as requested by the management. The management thoroughly examined the data and concluded it as a valid reflection of the business record. Table 1 provides the necessary data for developing a transportation model.

In Table 1, the supply and demand figures are in thousand (000) units and the upper left corner cells represent the unit transportation cost. A computer solution (*POM-OM for Windows* by Weiss, 2006) yields the optimal solution: Z (total cost) = \$2,630(000) $x_{11} = 10(000)$, $x_{14} = 30(000)$, $x_{23} = 30(000)$, $x_{25} = 20(000)$, $x_{31} = 10(000)$, $x_{32} = 50(000)$, $x_{41} = 20(000)$, $x_{45} = 30(000)$, and all other variables are zeros.

The dual of this problem is

$$\max \quad Z' = -40y_1 - 50y_2 - 60y_3 - 50y_4 + 40z_1 + 50z_2 + 30z_3 + 30z_4 + 50z_5$$

s.t.

$$\begin{aligned}
-y_1 + z_1 &\le 13 \\
-y_1 + z_2 &\le 18 \\
-y_1 + z_3 &\le 12 \\
-y_1 + z_4 &\le 15 \\
-y_1 + z_5 &\le 13 \\
-y_2 + z_1 &\le 15 \\
-y_2 + z_2 &\le 15 \\
-y_2 + z_3 &\le 10 \\
-y_2 + z_4 &\le 17 \\
-y_2 + z_5 &\le 10 \\
-y_3 + z_1 &\le 12 \\
-y_3 + z_2 &\le 12 \\
-y_3 + z_3 &\le 10 \\
-y_3 + z_4 &\le 16 \\
-y_3 + z_5 &\le 23 \\
-y_4 + z_1 &\le 19 \\
-y_4 + z_2 &\le 20 \\
-y_4 + z_3 &\le 17 \\
-y_4 + z_4 &\le 25 \\
-y_4 + z_5 &\le 15
\end{aligned}$$

$$y_1, y_2, \ldots, z_4, z_5 \ge 0$$

Since there are eight basic variables (i.e., $m+n-1 = 4+5-1 = 8$) in the primal optimal solution, the following relations can be found from Eq. (7):

Primal Variables	Dual Variables
$x_{11} = 10$	$z_1 = y_1+13$
$x_{14} = 30$	$z_4 = y_1+15$
$x_{23} = 30$	$z_3 = y_2+10$
$x_{25} = 20$	$z_5 = y_2+10$
$x_{31} = 10$	$z_1 = y_3+12$
$x_{32} = 50$	$z_2 = y_3+12$
$x_{41} = 20$	$z_1 = y_4+19$
$x_{45} = 30$	$z_5 = y_4+15$

In the dual problem, there are 9 equations in 20 unknowns.

Since y_i, $z_j \geq 0$, by setting one of the variables to zero (e.g., $y_4 = 0$), the following results are found, as shown in Table 2.

The implicit values of y_i represent the comparative advantage arising from a plant that is closer to the destination. Thus, the product per unit at plants A, B, and C is each worth \$6, \$5, and \$7 more, respectively, than the product at plant D. That is, plant D has the least advantage and $y_4 = 0$. The implicit values of z_j measure the delivered values at the destinations; the delivered value being equal to the value at the origin plant plus the unit transportation cost of the route used. For example, each unit of the product shipped from plant A to destination I is worth \$19 (= 6 + 13).

3.1. A Practical Technique of Generating the Dual Solution

The values of the dual variables (x_i, z_j) can be easily found in the following manner, without going through the mathematical procedures described previously. In Table 3, the primal variable solutions are represented by the circled numbers. Consider destinations I and V where there are multiple allocations exhibited by the circled numbers (i.e., basic variables). Since the delivered value at the destination I is the same whether the products are received from plants A, C, or D, the value at C has a \$1 advantage over that of A, and a \$6 advantage over that of D. That is, plant D has the least advantage among the three plants, A, C, and D; thus, we can assign an origin value of 0 to D. In destination V, there are two circled numbers, and plant B has a \$5 advantage over that of D. Thus, we obtain the same

Table 2. Dual Optimal Solution.

Plants	Value Per Unit	Destinations	Value Per Unit
A	$y_1 = 6$	I	$z_1 = 19$
B	$y_2 = 5$	II	$z_2 = 19$
C	$y_3 = 7$	III	$z_3 = 15$
D	$y_4 = 0$	IV	$z_4 = 21$
		V	$z_5 = 15$

Table 3. Primal Solution.

Dstn. / Plant	I	II	III	IV	V	Supply
A	13 (10)	18	12	15 (30)	13	40
B	15	15	10 (30)	17	10 (20)	50
C	12 (10)	12 (50)	10	16	23	60
D	19 (20)	20	17	25	15 (30)	50
Demand	40	50	30	30	50	200

solution, as shown in Table 2. This procedure is equally applicable to cases with unbalanced problems, as well as degeneracy and multiple optimal solutions.

3.2. Demand Forecasting for Facility Expansion/Contraction Decisions

The company in this study has the quarterly demand (sales) data dating back to 2001. Because of the seasonal changes in demand for farm and related equipments, quarterly demand forecasting is to be performed. Among the forecasting methods, the trend analysis (or multiple regression analysis) is often used for demand (sales) forecasting in transportation companies because of ease of data collection (or a lack of usable data).

In the trend analysis, the forecast equation can be expressed by

$$Y = \alpha + \beta\chi$$

where Y is the estimated demand, α the Y intercept, and β the parameter representing the average change in Y.

In the multiple regression analysis, the forecasted demand (a dependent variable) can be expressed by the factors (i.e., independent variables) affecting the demand volume as

$$Y = f(\chi_1,\ \chi_2,\ \chi_3,\ \chi_4 \cdots)$$

or

$$Y = \beta_0 + \beta_1\chi_1 + \beta_2\chi_2 + \beta_3\chi_3 + \beta_4\chi_4 + \varepsilon$$

where β_0 is the intercept, $\beta_1, \ldots, \beta_4$ the parameters representing the contributions of the factors, χ_1 the price of farm equipment, χ_2 the price of farm commodities, χ_3 the farm family income, χ_4 the acreage cultivated, and ε the error term.

A detailed description of the model development is beyond the scope of this study. A plethora of studies exists in the literature elsewhere. (For detailed analyses, see Kwak, Garrett, and Barone (1977) and Russell and Taylor (2009).) The linear trend analysis is adopted in this study because of inadequate data on independent variables (i.e., χ_1, χ_2, χ_3, χ_4) in the company record. In view of the seasonal variations in farm and related equipment sales, the forecasted Y is further adjusted by multiplying the seasonal index (k) as

$$Y_s = kY$$

where Y_s is the seasonally adjusted demand volume.

The seasonal index is derived using the ratio-to-trend method, as shown in Table 4.

For brevity, assume that the total demand has increased by 30 units. The company must expand plant C production facilities to accommodate the increase in demand, ceteris paribus (i.e., all other elements remain the same), because it has the best location advantage among the four plants. Likewise, if the total demand has decreased by 30 units, the plant D production will be reduced by 30 units accordingly, ceteris paribus. If the total demand volume has decreased by 50 units, plant D should be completely closed. The plant facility expansion/contraction (or overtime work/layoff) decisions may be easier when the changes in demand are known at a particular destination.

Table 4. Seasonal Variations: Ratio-to-Trend Method.

Quarter	Actual Value				Trend Value				Ratio = Actual/Trend			
	2001	2002	2003	2004	2001	2002	2003	2004	2001	2002	2003	2004
1	120.2	132.6	144.0	150.8	122.2	134.5	146.8	159.1	.984	.986	.981	.948
2	126.1	141.3	158.9	171.6	125.3	137.6	149.9	162.1	1.006	1.027	1.060	1.059
3	133.3	147.8	163.1	178.4	128.4	140.6	152.9	165.2	1.038	1.051	1.067	1.080
4	130.2	140.1	151.0	173.9	131.4	143.7	156.0	168.3	.991	.975	.968	1.033

Quarter	Actual Value			Trend Value			Ratio = Actual/Trend			Mean	Seasonal Index
	2005	2006	2007	2005	2006	2007	2005	2006	2007		
1	162.6	175.0	180.9	171.3	183.6	195.9	.947	.953	.923	.961	.957
2	179.0	190.8	206.4	174.4	186.7	198.7	1.026	1.022	1.039	1.034	1.029
3	184.5	186.6	210.7	177.5	189.8	202.0	1.039	.983	1.043	1.043	1.038
4	180.7	179.4	196.7	180.6	192.6	205.0	1.001	.931	.960	.980	.976
										4.018	4.000

Note: $Y = 165.2 + 3.07\chi$ (origin: Quarter 2, 2004; χ: unit, Quarter; Y: average quarterly sales (000)). Seasonal adjustment factor: $k = 4.000/4.018 = 0.996$.

4. CONCLUDING REMARKS

The dual transportation problem was analyzed for plant facility expansion/contraction decision making in manufacturing operations, along with demand forecasting with an illustrative case example. This duality analysis of transportation problem is equally applicable to other resource allocations problems in business and industry. In manufacturing operations, the firm's accounting personnel can be assigned to a variety of jobs (e.g., accounts receivables, accounts payable, sales auditing, payroll) when the problem is formulated and presented in transportation matrix format. The qualifications of the accounting personnel can be implicitly rated for effective control and better operations management.

REFERENCES

Badri, M. A. (1999). Combining the analytic hierarchy process and goal programming for global facility-allocation problem. *International Journal of Production Economics*, *62*, 237–248.

Bowen, W. M. (1995). A Thurstonian comparison of the analytic hierarchy process and probabilistic multidimensional scaling through application to the nuclear waste site selection decision. *Socio-economic Planning Science*, *29*, 151–164.

Campbell, J. F., Ernst, A. T., & Krishnamoorthy, M. (2005). Hub arc location problems: Part 1 – Introduction and results. *Management Science*, *51*, 1540–1555.

Current, J. H., Min, H., & Schilling, D. (1990). Multiobjective analysis of facility location decisions. *European Journal of Operational Research*, *49*, 295–308.

Drezner, Z., & Hamacher, H. (2004). *Facility location: Application and theory*. Berlin, Germany: Verlag.

Erlenkotten, D. (1978). A dual-based procedure for uncapacitated facility location. *Operations Research*, *26*, 992–1009.

Farahani, R. Z., & Asgari, N. (2007). Combination of MCDM and converging techniques in a hierarchical model for facility location: A case study. *European Journal of Operational Research*, *176*, 1839–1858.

Glover, F. (1990). A Tabu search: A tutorial. *Interfaces*, *20*, 74–94.

Ho, W., Lee, C. K. M., & Ho, G. T. S. (2008). Optimization of the facility location-allocation problem in a customer-driven supply chain. *Operations Management Research*, *1*, 69–79.

Jaramillio, J. H., Bhadury, J., & Batta, R. (2002). On the use of generic algorithms to solve location problems. *Computers & Operations Research*, *29*, 761–779.

Keeney, R. L. (1994). Using values in operations research. *Operations Research*, *42*, 793–813.

Klamroth, K. (2002). *Single facility location problems with barriers*. Berlin, Germany: Springer Verlag.

Klorpela, J., & Truominen, M. (1996). A decision aid in warehouse site selection. *International Journal of Production Economics*, *45*, 169–181.

Korhonen, P., & Syrjanen, M. (2004). Resource allocation based on efficiency analysis. *Management Science*, *50*, 1134–1144.

Kwak, N. K. (1973). *Mathematical programming with business applications*. New York: McGraw-Hill.

Kwak, N. K., Garrett, W. A., & Barone, S. (1977). Stochastic model of demand forecasting for technical manpower. *Management Science*, *23*, 1089–1098.

Kwak, N. K., & Schniederjans, M. J. (1985). A goal programming model as an aid to facility location analysis. *Computers & Operations Research*, *12*, 151–161.

Mehrez, A., Sinuany-Stern, Z., Arad-Geva, T., & Binyamin, S. (1996). On the implementation of quantitative facility location models: The case of a hospital in a rural region. *Journal of the Operational Research Society*, *47*, 612–626.

Quarterman, L. (1998). Points to consider in selecting facilities planning software. *ITE Solutions*, *30*, 42049.

Revelle, C. S., & Laporte, G. (1996). The plant location plan: New models and research prospects. *Operations Research*, *44*, 864–874.

Russell, R. S., & Taylor, B. W., III. (2009). *Operations management*. New York: Wiley.

Saaty, T. L. (1980). *The analytic hierarchy process*. New York: McGraw-Hill.

Schniederjans, M. J. (1999). *International facility acquisition and location analysis*. Westport, CT: Quorum Books.

Sinha, S. B., & Sastry, S. V. C. (1987). A goal programming model to resolve a site location problem. *Interfaces*, *12*, 251–256.

Sule, D. R. (1988). *Manufacturing facilities: Location, planning, and design*. Boston, MA: PWS-Kent Publishing.

Tavana, M., & Banerjee, S. (1995). Evaluating strategic alternatives: An analytic hierarchy process. *Computers & Operations Research*, *22*, 731–743.

Weiss, H. J. (2006). *POM-QM for windows (Version 3)*. Upper Saddle River, NJ: Pearson/ Prentice Hall.

MAKE-TO-ORDER PRODUCT DEMAND FORECASTING: EXPONENTIAL SMOOTHING MODELS WITH NEURAL NETWORK CORRECTION

Mark T. Leung, Rolando Quintana and An-Sing Chen

ABSTRACT

Demand forecasting has long been an imperative tenet in production planning especially in a make-to-order environment where a typical manufacturer has to balance the issues of holding excessive safety stocks and experiencing possible stockout. Many studies provide pragmatic paradigms to generate demand forecasts (mainly based on smoothing forecasting models.) At the same time, artificial neural networks (ANNs) have been emerging as alternatives. In this chapter, we propose a two-stage forecasting approach, which combines the strengths of a neural network with a more conventional exponential smoothing model. In the first stage of this approach, a smoothing model estimates the series of demand forecasts. In the second stage, general regression neural network (GRNN) is applied to learn and then correct the errors of estimates. Our empirical study evaluates the use of different static and dynamic

Advances in Business and Management Forecasting, Volume 6, 249–266

ISSN: 1477-4070/doi:10.1108/S1477-4070(2009)0000006015

smoothing models and calibrates their synergies with GRNN. Various statistical tests are performed to compare the performances of the two-stage models (with error correction by neural network) and those of the original single-stage models (without error-correction by neural network). Comparisons with the single-stage GRNN are also included. Statistical results show that neural network correction leads to improvements to the forecasts made by all examined smoothing models and can outperform the single-stage GRNN in most cases. Relative performances at different levels of demand lumpiness are also examined.

1. INTRODUCTION

Demand forecasting has long been an imperative tenet in production planning. In a make-to-order environment where production schedules generally follow the fulfillment of orders, a typical manufacturer still has to balance the paradoxical issues of producing too much which leads to excessive inventory and producing inadequately which causes backlog. Either of the two scenarios can often result in higher cost and more waste. Hence, it would be advantageous for a manufacturer to obtain accurate estimates of demand orders (volumes) even before the orders are actually received. In light of this practical need, many studies have formulated pragmatic paradigms to generate demand forecasts using a wide spectrum of methods and models. Nevertheless, it seems that most of the conventional production forecasting models adopted mainly align with the stream of exponential smoothing forecasting. Over the past decade, more innovative models have been emerging as alternatives due to the advancement of computational intelligence. One of such stream of research is the use of artificial neural networks (ANNs), which has been applied to solve problems encountered in different manufacturing settings including production scheduling, cell design and formation, quality control, and cost estimation.

In this study, we apply ANN to forecast demand orders based on historical (observable) data as well as to improve the forecasts made by other models. In other words, ANN is used to correct the errors of a base model's forecasts in a two-stage adaptive forecasting framework. Conventional exponential smoothing models are chosen and facilitate as the basis models in this forecasting research because of their popularity among industrial practitioners and academic circle. Because of the existence of different approaches and adaptations of the smoothing concept, we

explicitly consider two groups of smoothing models – one that is based on static model parameters, which do not autonomously change over time and the other that utilizes dynamically self-adjusted smoothing control constants. The static models are simpler but may be subject to a slower response rate. On the other hand, the more mathematically complex dynamic models should create better forecasts. With the current computer technology, these models can be easily implemented.

Thus, the demand forecasting study consists of single- and two-stage models. The single-stage models include static and dynamic smoothing models and a neural network, namely, general regression neural network (GRNN). Each of the two-stage models is made up of a smoothing model and GRNN. Essentially, a smoothing model is used as a basis and forms the series of demand estimates in the first stage of forecasting. Consequently, GRNN is employed to "learn" the error residuals of these estimates and makes proper adjustments (corrections) to the estimates. Our conjecture is that the two-stage forecasting models with error correction capability should perform better than the single-stage models. Another purpose of this chapter is to show that the use of neural network in error correction in this two-stage framework can lead to improvement in the original forecasts even generated by the neural network itself. It is because neural network mostly suffers the issue of unexplored (untrained) state space due to its weakness in extrapolation. An exponential smoothing model, which is capable of extrapolating into unknown state space, can alleviate this weakness. Hence, combining a smoothing model with a neural network may create synergy in forecasting. All single- and two-stage models are evaluated according to a spectrum of measures such as root mean squared error (RMSE), improvement ratios (IRs) over the base model and the single-stage neural network, information contents, forecast bias and proportion. The empirical analysis also evaluates the relative performances of models at different levels of demand lumpiness or volatility. It is believed that the higher the lumpiness, the more uncertainty associated with demand orders and thereby the more deterioration to a model's forecasting capacity.

The chapter is organized as follows. In Section 2, a conceptual background of the single-stage forecasting models is briefly summarized. This includes three conventional static and two dynamically self-adjusted smoothing models, as well as GRNN, the neural network employed in the second stage of the two-stage forecasting framework. In Section 3, the methodologies for both single- and two-stage forecasting are explained. The section also describes the data set and the horizons for model estimation and performance evaluation. Moreover, criteria for data categorization with respect to demand

lumpiness are outlined. In Section 4, results from the empirical investigation are presented and discussed. Section 5 concludes the chapter.

2. BACKGROUND AND BASIC METHODOLOGIES

This study compares the lumpy demand forecasting capabilities of an array of exponential smoothing models with that of a neural network. The study also attempts to calibrate any possible synergetic effect on these smoothing models due to error corrections performed by a neural network within a two-stage forecasting framework. In other words, the empirical experiment evaluates the degrees of enhancement on traditional demand forecasts subject to error corrections by GRNN. Since different types of exponential smoothing models exist, we select the first three models based on their wide popularity in industrial practice. Nonetheless, their model parameters are fixed (static) and do not adapt spontaneously to changes in demand. On the other hand, the fourth and the fifth chosen exponential smoothing models consist of dynamically updated smoothing parameters and are thus capable of adjusting their values autonomously. A brief exposition of these five exponential smoothing models is provided in the following sections.

2.1. Simple Exponential Smoothing Model

In our empirical experiment, we examine three types of static exponential smoothing models. Generally speaking, they are the time series methods most commonly used in demand forecasting in the industry and widely embraced in academic textbooks in the field of operations management. Essentially, these methods smooth out previously observed demands and continuously generate forecasts by incorporating more recent historical information into the previous estimations. In other words, the concept is based on averaging past values of historical demand data series in a decreasing exponential manner. The historical demands are weighted, with larger weight given to the more recent data.

As the basic building block of most exponential smoothing forecasting, a simple exponential smoothing model can be written as

$$F_t = \alpha A_t + (1 - \alpha) F_{t-1} \tag{1}$$

where F_t is the forecasted demand for period $(t+1)$ made in period t, A_t the actual demand observed in period t, F_{t-1} the previous demand forecast for

period t and is made in period $(t-1)$, and α the smoothing constant (where $0 \leq \alpha \leq 1$), which does not change over time. The above model suggests that the smoothing procedure has the capacity of feeding back the forecast error to the system and correcting the previous smoothed (forecasted) value. For a more detailed explanation, readers can refer to Anderson, Sweeney, and Williams (2005) and Makridakis and Wheelwright (1977).

2.2. Holt's Exponential Smoothing Model

An issue about the simple exponential smoothing model is that its estimates will lag behind a steadily rising or declining trend. In light of this, Holt (1957) developed a linear exponential smoothing model with trend adjustment. The model involves two iterative estimations, one for the nominal smoothed value and the other for the trend adjustment. Technically, each of these estimations is treated as a separate exponential smoothing and requires its own smoothing constant. The two-parameter forecasting system can be expressed by the following system of equations:

$$S_t = \alpha(A_t) + (1 - \alpha)(S_{t-1} + T_{t-1}) \tag{2}$$

$$T_t = \beta(S_t - S_{t-1}) + (1 - \beta)T_{t-1} \tag{3}$$

$$F_t = S_t + T_t \tag{4}$$

where S_t is the nominal forecast made in period t for period $(t+1)$, T_t the trend forecast made in period t for period $(t+1)$, A_t the actual demand observed in period t, α the nominal smoothing constant $(0 \leq \alpha \leq 1)$; and β the trend smoothing constant $(0 \leq \beta \leq 1)$.

The first equation is similar to the static constant exponential smoothing except a trend estimate is appended for adjustment of the previous demand forecast. The output constitutes an estimate for the nominal smoothed value. The second equation is used to compute the trend estimate in the first equation. This is done by taking a weighted average of the previous trend estimate and the difference between successive nominal smoothed values. In the third equation, the nominal smoothed value is combined with the trend estimate to form the demand forecast for the next period. Holt's model requires the use of two static parameters, α and β, in the estimations of smoothed and trend values, respectively. In our empirical experiment, their values are determined jointly based on the demand pattern exhibited during the in-sample period.

2.3. Winter's Exponential Smoothing Model

Although the Holt's model explicitly considers the trend of demand by separating it from the general (nominal) forecast of demand, the model itself can be further improved by taking into account of possible seasonal effect, that is, cyclical upward and downward movements over a relatively longer period (e.g., a year) than the time frame of each forecast period (e.g., a week). It should be noted that seasonality is simply a generic descriptor to denote cyclical or repetitive demand patterns.

By extending Holt's three-equation model, Winter (1960) developed an exponential smoothing with trend and seasonal components. The model contains a set of four equations. The conceptual background lies on the notion that a forecast can be divided into three components – the nominal forecast, the trend forecast, and the seasonal forecast. Hence, we try to estimate these three components separately. After all three estimates have been made, they are combined to form an aggregate forecast for demand. However, the way to combine the component forecasts is different from the way as in the Holt's exponential smoothing model.

The exponential smoothing model with trend and seasonal components is represented by the following iterative equations:

$$S_t = \alpha\left(\frac{A_t}{I_{t-L}}\right) + (1-\alpha)(S_{t-1} + T_{t-1}) \tag{5}$$

$$T_t = \beta(S_t - S_{t-1}) + (1-\beta)T_{t-1} \tag{6}$$

$$I_t = \gamma\left(\frac{A_t}{S_t}\right) + (1-\gamma)I_{t-1} \tag{7}$$

$$F_t = (S_t + T_t)I_{t-L+1} \tag{8}$$

where S_t is the nominal forecast made in period t for period $(t+1)$, T_t the trend forecast made in period t for period $(t+1)$, I_t the seasonal index used in period t to adjust the forecast for period $(t+1)$, A_t the actual demand observed in period t, L the number of periods in a typical cycle of demand movements, α the nominal smoothing constant $(0 \le \alpha \le 1)$, β the trend smoothing constant $(0 \le \beta \le 1)$, and γ the seasonal smoothing constant $(0 \le \gamma \le 1)$.

2.4. Adaptive Exponential Smoothing Model

Makridakis, Wheelwright, and McGee (1983) and Mabert (1978) described an extension to traditional static exponential smoothing models, generally known as adaptive exponential smoothing. This approach continuously evaluates the performance in the previous period and updates the smoothing constant. The form of the adaptive exponential smoothing model is a modification and extension to that of the simple exponential smoothing model with static smoothing constant

$$F_{t+1} = \alpha_t A_t + (1 - \alpha_t) F_t \tag{9}$$

$$\alpha_{t+1} = \left| \frac{E_t}{M_t} \right| \tag{10}$$

$$E_t = \beta e_t + (1 - \beta) E_{t-1} \tag{11}$$

$$M_t = \beta |e_t| + (1 - \beta) M_{t-1} \tag{12}$$

$$e_t = A_t - F_t \tag{13}$$

where F_t is the forecast for period t, A_t the actual demand observed in period t, α and β are model parameters between 0 and 1, and $|\bullet|$ denotes absolute value. It should be pointed out that α_t is a dynamic smoothing constant with its value updated in each period. β can be viewed as a control parameter to the responsiveness of the dynamic smoothing constant (α_t) to demand changes. In summary, the interconnected iterative system of equations provides feedback to both demand estimation and updates the value of the smoothing constant based on the observed changes in recent historical demands.

2.5. Dynamic Exponential Smoothing Model Using Kalman Filter

Although the adaptive exponential smoothing model dynamically updates the smoothing constant, the issues of selecting the control parameter for responsiveness (β) and choosing an initial value of (α_t) remain. To resolve these issues, Quintana and Leung (2007) presented a dynamic exponential smoothing model with Kalman filter. Essentially, the Kalman filter adopted in the forecasting paradigm calibrates demand observations to estimate the state of a linear system and utilizes knowledge from states of measurements

and system dynamics. Technically speaking, at any current period j, the Kalman filter weighting function $W(j+1)$ developed as a smoothing variable (α) for forecasting lumpy demand at period ($j+1$) can be expressed by the following mathematical structure:

$$\alpha = W(j+1) = \left\{ \frac{\left[\frac{\sum_{i=j-N+1}^{j} (D_i - \bar{D})^2}{N-1} \right]^{1/2}}{\left| \left[\frac{\sum_{i=j-N+1}^{j-1} (D_i - \bar{D})^2}{N-1} \right]^{1/2} - \left[\frac{\sum_{i=j-N+1}^{j} (D_i - \bar{D})^2}{N-1} \right]^{1/2} \right|} \right\} \tag{14}$$

where W is the weighting function (adaptive smoothing variable), j the current period, N the maximum number of past periods used, and D the demand.

The numerator is the standard deviation of the demand from the current period back N periods, whereas the denominator is the difference between the standard deviations for N previous periods from the current and the previous periods, respectively. In this manner, weighting function acts as an estimation regulator in that it will dampen the effects of statistical outliers. For a more detailed exposition of the methodology, readers should refer to Quintana and Leung (2007).

2.6. General Regression Neural Network

GRNN is a form of ANNs first proposed by Specht (1991). It is a multilayer feed forward-learning network capable of approximating the implied relationship from historical data. Also, it has the distinctive features of swift learning, requiring only a single pass in training paradigm, and being insensitive to infrequent outliers (given the training data set is sufficiently large).

Essentially, GRNN is able to estimate any arbitrary relationship between a given set of input variables and its corresponding outputs. This estimation procedure is carried out by the network during the training process. On the completion of training, the deduced relationship is used to compute the (expected value of) output vector based on a given input vector. In the GRNN model, estimation of a dependent variable y with respect to a given

vector of independent variables **X** can be regarded as finding the expected value of y conditional on the value of **X**. The following equation summarizes this statistical concept:

$$E[y|\mathbf{X}] = \frac{\int_{-\infty}^{\infty} yf(\mathbf{X}, y)dy}{\int_{-\infty}^{\infty} f(\mathbf{X}, y)dy} \tag{15}$$

where y is the output value estimated by GRNN, **X** the input vector for the estimation of y; and $f(\mathbf{X}, y)$ the joint probability density function of **X** and y learned by GRNN from the available training data set.

Justifications for the choice and use of GRNN architectural design for neural network forecasting in this study are primarily due to its relative simplicity in training and its rather encouraging results and stable performances found by other studies. For the sake of brevity, readers can refer to Wasserman (1993) for a complete explanation of the foundation and operational logic of this specific design of neural network model.

3. FORECASTING DEMAND

3.1. Data and Single-Stage Forecasting of Demand

Our data set is based on an industrial consulting project with a Mexican production facility supplying parts to major automobile manufacturers. Demand order information of more than 400 SKUs was obtained from the management. For the sake of a more focused experiment, only observations of the items with the 10 largest aggregate demand volumes in each lumpiness category are used in our comparative evaluation. There are three categories of lumpiness, representing different levels of demand order volatility. To classify the level of lumpiness, the manufacturing company defines a "low" lumpy environment as one within ± 1 standard deviation from the mean demand. Medium and high lumpiness are defined as within ± 2 and beyond ± 2 standard deviations from the mean, respectively.

The provided weekly demand data run from January 1997 to December 2005.[1] In our empirical experiment, the historical data series is divided into two sample periods – the estimation (in-sample) and the test (out-of-sample) periods. The estimation period covers observations from January 1997 to December 2002 and is used for establishment of smoothing parameters in various single-stage (fixed constant, Holt's, Winter's, and Kalman filter) models. It also serves as the training period for the single-stage GRNN

forecasting and the two-stage GRNN adaptive error correction. Moreover, the first year in the estimation period is reserved as an initialization period for estimations by the Holt's smoothing, the Winter's smoothing, and both single- and two-stage GRNN models. On the basis of an assessment of performances in the estimation period, the specification of each model type is selected and subject to out-of-sample testing. The three-year test period goes from January 2003 to December 2005 and is reserved strictly for the purpose of performance evaluation.

3.2. Two-Stage Demand Forecasting with Error Correction

Given its demonstrated performance, the two-stage error correction framework described by Chen and Leung (2004) for foreign exchange forecasting is modified and adapted to our problem environment. For the two-stage demand forecasting, a smoothing model is estimated and then its forecasts are subsequently corrected by GRNN. In the first stage, each of the five static and dynamic exponential smoothing models is estimated based on the paradigm described in the previous section. After that, residuals for the in-sample forecasts from January 1998[2] to December 2002 are computed. GRNN is applied to estimate the error distribution. As we move forward into the out-of-sample period, new forecasts are generated from the smoothing model and new residuals are produced. Hence, as the data of a week become observable, the residual associated with that week can be generated by subtracting the demand forecast from the newly observed demand, which is now observable. The training set is then updated by incorporating this newly computed residual and eliminating the oldest residual observation. Then, GRNN is retrained using the updated residual series. The forecast for the expected residual of following week is then generated using the retrained GRNN. An error-corrected demand forecast for the following week can be attained by adding the following week's forecasted residual to the original single-stage forecast computed by the smoothing model. This two-stage forecasting paradigm is repeated for the five smoothing models.

4. RESULTS

Out-of-sample performances of the forecasting models in our empirical study are tabulated in Table 1. The results with respect to RMSE are

Table 1. Out-of-Sample Comparison of Root Mean Squared Errors Among Various Exponential Smoothing Models and Performance Improvements by Adaptive Neural Network Correction.

Model	Root Mean Squared Error (RMSE)	Improvement Over the Original Single-Stage Smoothing (%)	Improvement Over the Single-Stage GRNN (%)
Single-stage smoothing and GRNN models			
Simple ES	68.95		
Holt ES	65.79		
Winter ES	63.04		
Adaptive ES	59.49		
Kalman ES	58.72		
GRNN	56.83		
Two-stage smoothing models with GRNN correction			
Simple-GRNN	61.59	10.67	−8.38
Holt-GRNN	56.57	14.01	0.46
Winter-GRNN	55.40	12.12	2.52
Adaptive-GRNN	50.62	14.91	10.93
Kalman-GRNN	51.39	12.48	9.57

Note: Two-stage adaptive exponential smoothing with GRNN correction yields the minimum RMSE among all models. All two-stage models with neural network correction gain significant performance relative to their original smoothing models. Four smoothing models – Holt, Winter, Adaptive, and Kalman filter, in conjunction with GRNN correction outperforms the single-stage GRNN. RMSE improvement ratio (IR) is computed as

$$\mathrm{IR} = \frac{\mathrm{RMSE}_2 - \mathrm{RMSE}_1}{\mathrm{RMSE}_1}$$

where RMSE_1 is the root mean squared error of the forecasts made by original single-stage smoothing model or the single-stage GRNN, whereas RMSE_2 is the root mean squared error of forecasts estimated by the corresponding two-stage model with neural network correction.

compared within the groups of single- and two-stage models. For the single-stage category of models, it can be seen that dynamically adjusted exponential smoothing models are better than the more conventional smoothing models with static constants. Also, the neural network model (GRNN) outperforms the two dynamic smoothing models, both of which yield pretty close RMSEs. For the two-stage models, the RMSEs of the two dynamic smoothing models in conjunction with GRNN are lower than their counterparts based on static smoothing models with neural network correction. In summary, the results support the conjecture that smoothing models with dynamic adjustment capability are generally more accurate than the conventional static smoothing models as observed in our manufacturing order data set.

Table 1 also compares two- and single-stage models by evaluating the improvement of GRNN error correction used in a two-stage model over its single-stage counterpart without error correction. Specifically, the IR is computed by the difference between the RMSE of a two-stage model and that of its single-stage counterpart without correction divided by the RMSE of the single-stage counterpart without correction. This computation can be expressed as follow:

$$\mathrm{IR} = \frac{\mathrm{RMSE}_2 - \mathrm{RMSE}_1}{\mathrm{RMSE}_1} \tag{16}$$

where RMSE_1 is the RMSE of the forecasts made by original single-stage smoothing model or the single-stage GRNN, whereas RMSE_2 is the RMSE of forecasts estimated by the corresponding two-stage model with neural network correction.

As shown in Table 1, the computed ratios indicate that a minimum of 10% improvement can be obtained across all smoothing models when neural network error correction is used. This finding reveals the synergetic effect of combining a smoothing model with neural network in demand forecasting. In addition, improvement over the single-stage GRNN is also computed for each two-stage smoothing model. Unlike their improvements over the original single-stage smoothing models, the improvements over single-stage GRNN may or may not be significantly greater than zero. Nonetheless, an examination of the RMSEs suggests that significant improvements over GRNN are attained in the cases of dynamic smoothing models with neural network correction. This observation may be attributed to the excellent performance of the single-stage GRNN model in the forecasting of the out-of-sample demand series. Essentially, the better forecasts from GRNN make the poorly performed two-stage models based on static constants more difficult to catch up even neural network error correction is adopted. Among all tested models, adaptive exponential smoothing model with GRNN correction leads to the best set of out-of-sample forecasts (with $\mathrm{RMSE} = 50.62$.) Besides, this two-stage model captures the largest improvement relative to both the original single-stage adaptive exponential model and the single-stage GRNN.

In light of these findings, we conduct informational content tests to cross-examine and validate the better performances induced by neural network error correction. The informational content test developed by Fair and

Shiller (1990) involves running regressions of the realized correlation on a constant and a pair of demand forecasts. The regression equation is

$$Z_{t+1} = \alpha + \beta Z^e_{1t,t+1} + \gamma Z^e_{2t,t+1} + \mu_t \tag{17}$$

where $Z^e_{1t,t+1}$is the one-week ahead forecast made by model 1 at time t, and $Z^e_{2t,t+1}$is the one-week ahead forecast made by model 2 at time t. In addition, because of potential multicollinearity, Wald tests are performed on two possible restrictions: first, that the coefficient on the benchmark model (model 1) is equal to zero; and second, that the coefficient on the error correction model being tested (model 2) is equal to zero. Wald test statistic is based on χ^2 distribution. For methodological details of the empirical test, readers can refer to the original article.

Table 2 reports the results of the informational content tests. In panel A, the single-stage GRNN is compared with its two-stage counterparts with adaptive error correction by GRNN. According to the Wald tests, all two-stage models with GRNN correction are significant at the 10% level, indicating that the out-of-sample forecasts from each of these two-stage models contain information not revealed in the single-stage GRNN forecasts. On the contrary, the vice versa is not correct, that is, forecasts from the single-stage GRNN model does not contain additional information beyond the forecasts generated by the two-stage models. Since all two-stage models involve GRNN correction in the second stage, the results from the informational content tests show the usefulness of error correction and the capacity of neural network on the analysis (and prediction) of demand residuals. Further, this observation is possibly a consequence of the weakness in extrapolation commonly associated with neural network forecasting.

In Table 1, we conclude that the adaptive exponential smoothing model with GRNN correction yields the best result among the demand forecasting models in the study. Hence, it is logical to compare its performance with those of the other two-stage correction models using the informational content tests. Panel B (Table 2) points out mixed results based on the Wald tests. Adaptive exponential smoothing with correction generates forecasts with information not contained in the forecasts from the two-stage models built on simple and Holt's exponential smoothing. However, its forecasts do not have the informational advantage over the two-stage models based on the Winter's and the Kalman filter exponential smoothing. Besides, the forecasts from the Winter's smoothing model with correction demonstrates

Table 2. Informational Content Tests of Alternative Forecasts during the Out-of-Sample Period.

Panel A: Comparisons of single-stage GRNN with smoothing models with GRNN correction

Constant	Single-Stage GRNN	Simple-GRNN	Holt-GRNN	Winter-GRNN	Adaptive-GRNN	Kalman-GRNN	χ_1^2	χ_2^2
0.0103	0.4962	0.6938					1.720	2.227[a]
−0.0236	0.3822		0.6645				1.517	1.925[a]
−0.0279	0.4160			0.7327			1.312	2.664[a]
−0.0564	0.3551				0.9358		1.047	3.342[a]
−0.0515	0.3749					0.9130	1.038	3.183[a]

Panel B: Comparisons of adaptive ES with GRNN correction with other smoothing models with GRNN correction

Constant	Adaptive-GRNN	Simple-GRNN	Holt-GRNN	Winter-GRNN	Kalman-GRNN	χ_1^2	χ_2^2
0.0520	0.6846	0.2495				2.385[a]	0.803
0.0380	0.5970		0.3613			2.198[a]	1.412
−0.0335	0.4068			0.5868		1.301	2.086[a]
0.0375	0.2593				0.2257	0.814	0.792

Note: The informational content test involves running regressions of the actual demand on a constant and a pair of demand forecasts. The regression equation is

$$Z_{t+1} = \alpha + \beta Z^e_{1t,t+1} + \gamma Z^e_{2t,t+1} + \mu_t$$

where $Z^e_{1t,t+1}$ is the one-week ahead forecast made by model 1 at time t, and $Z^e_{2t,t+1}$ the one-week ahead forecast made by model 2 at time t. The first χ^2 column corresponds to the test statistic from Wald test (distributed as χ^2) on the restriction that the coefficient on model 1 forecasts is equal to zero. The second χ^2 column corresponds to the test statistic from Wald test on the restriction that the coefficient on model 2 forecasts is equal to zero. Simple-GRNN is based on a fixed smoothing constant of 0.62.

[a]Indicates that the regression coefficient is different from zero at the 10% significance level according to the Wald χ^2 test statistic.

information content not found in the forecasts estimated by the adaptive-GRNN model, the best model in terms of RMSE. As a concluding remark, although our experimental results do not identify the definitely best performer among the two-stage forecasting models, the findings do provide evidence of the value of neural network correction for improving the accuracy of demand forecasts. Furthermore, the findings give some guidance to the selection of demand forecasting model in the future.

Encouraged by the effectiveness of neural network correction in exponential smoothing forecasting of demand, we compare the relative forecasting strengths of various models with respect to the levels of demand lumpiness, which have already been defined and explained in "Data and Single-Stage Forecasting of Demand" section. Specifically, we perform Theil's (1966) decomposition tests to examine the characteristics of the out-of-sample forecasts estimated by the models. The Theil's decomposition test is conducted by regressing the actual observed demand on a constant and the demand forecast estimated by a particular model

$$A_t = \alpha + \beta D_t + \varepsilon_t \tag{18}$$

where A_t is the actual demand at period t, D_t the forecasted demand for period t made at period $t-1$, and ε_t the error term. The constant α (bias coefficient) should be insignificantly different from zero and the coefficient β for estimated demand (regression proportion coefficient) should be insignificantly different from one for the forecast to be acceptable.

Results of the decomposition test are displayed in Table 3. It should be noted that the superior a and b denote that the bias coefficient (α) and the proportion coefficient (β) are *insignificantly* different from 0 and 1, respectively. At low lumpiness level, the two-stage models with GRNN correction perform well in terms of insignificant bias ($\alpha = 0$) and parallel proportion to actual demand ($\beta = 1$). This conclusion still holds for forecasting demand at the moderate lumpiness level except for the two-stage model based on simple exponential smoothing. However, when demand becomes highly lumpy or volatile, only a few models generate good forecasts. GRNN is the only single-stage model, which yields forecasts with a bias insignificantly different from zero and, at the same time, a proportion coefficient insignificantly different from one. For the two-stage models, dynamic adaptive and Kalman filter exponential smoothing with GRNN are the two constructs satisfying the evaluation criteria. Performances of two-stage models built on static smoothing models drastically deteriorate when demand becomes highly lumpy. It is suspected that the dynamically adjusted smoothing constants adopted in these two models instigate swift adaptation and thus are better coping with more uncertain demand fluctuations during volatile periods. Again, the results echo the superior performances of adaptive-GRNN and Kalman-GRNN reported in previous tables.

Table 3. Theil's Decomposition Test Results for Various Forecasting Models at Different Levels of Demand Lumpiness during the Out-of-Sample Period.

Model	α (Bias Coefficient)	$t(\alpha = 0)$	β (Proportion Coefficient)	$t(\beta = 1)$
Low lumpiness				
Simple ES	3.38	2.95	2.71	3.15
Holt ES	2.64	1.57[a]	1.86	2.02
Winter ES	3.45	3.12	1.61	1.76[b]
Adaptive ES	2.26	1.35[a]	1.48	1.31[b]
Kalman ES	2.43	1.48[a]	1.39	1.17[b]
GRNN	−1.63	−1.05[a]	0.83	−0.53[b]
Simple-GRNN	2.58	1.62[a]	2.12	2.32
Holt-GRNN	1.52	0.97[a]	0.80	−0.54[b]
Winter-GRNN	−1.20	−0.73[a]	0.84	−0.52[b]
Adaptive-GRNN	1.12	0.69[a]	0.85	−0.52[b]
Kalman-GRNN	0.98	0.56[a]	0.89	−0.50[b]
Moderate lumpiness				
Simple ES	7.38	3.37	4.31	3.56
Holt ES	4.93	2.13	2.53	2.64
Winter ES	4.18	1.79[a]	2.04	2.02
Adaptive ES	3.85	1.61[a]	1.79	1.72[b]
Kalman ES	3.46	1.45[a]	1.67	1.65[b]
GRNN	2.77	1.18[a]	0.80	−0.68[b]
Simple-GRNN	5.08	2.33	2.62	2.78
Holt-GRNN	4.05	1.71[a]	0.67	−1.06[b]
Winter-GRNN	−3.30	−1.37[a]	0.76	−0.87[b]
Adaptive-GRNN	2.72	1.17[a]	0.86	−0.55[b]
Kalman-GRNN	2.58	1.08[a]	0.84	−0.55[b]
High lumpiness				
Simple ES	25.74	3.68	4.73	3.72
Holt ES	19.03	3.04	3.57	3.04
Winter ES	15.76	2.67	3.19	2.83
Adaptive ES	11.68	2.07	2.16	2.20
Kalman ES	11.43	2.04	2.37	2.28
GRNN	−10.63	−1.83[a]	0.51	−1.47[b]
Simple-GRNN	17.08	2.78	3.62	3.13
Holt-GRNN	15.43	2.58	1.89	1.95[b]
Winter-GRNN	13.08	2.33	1.62	1.59[b]
Adaptive-GRNN	9.15	1.53[a]	0.69	−1.31[b]
Kalman-GRNN	9.68	1.64[a]	0.56	−1.40[b]

Note: The Theil's decomposition test is specified as follow:

$$A_t = \alpha + \beta D_t + \varepsilon_t$$

where A_t is the actual demand at period t, D_t the forecasted demand for period t made at period $t-1$, and ε_t the error term.

[a] t values indicate that the null hypothesis of H_0: $\alpha = 0$ cannot be rejected at the 5% significance level.

[b] t values indicate that the null hypothesis of H_0: $\beta = 1$ cannot be rejected at the 5% significance level.

5. CONCLUSIONS

In this chapter, we compare the lumpy demand forecasting capabilities of an array of exponential smoothing models with that of GRNN. The study also attempts to calibrate any possible synergetic effect on these smoothing models due to error corrections performed by a neural network within a two-stage forecasting framework. In other words, our empirical experiment evaluates the degrees of enhancement on traditional demand forecasts subject to error corrections by GRNN. The exponential smoothing models considered in this study belong to two types, static models with constant parameters and dynamic models with self-adjusted smoothing constants. This array of five smoothing models serves as the basis of the two-stage forecasting framework and creates the first-cut demand estimates. In the second stage, these forecasts are corrected by the error residuals estimated by GRNN.

Results of the experiment indicate that forecasting accuracy of all (static and dynamic) smoothing models can be improved by GRNN correction. This is a supporting evidence of the synergy realized by combining the capacity of conventional forecasting model with neural network. Results also reveal that two-stage models probably perform better than just the single-stage GRNN. In addition, the study explores the overlapping of information contents between single-stage GRNN and two-stage models with GRNN correction. It is shown that the forecasts from all two-stage models possess information not revealed in the single-stage GRNN. This observation is possibly a consequence of the weakness in extrapolation commonly associated with neural network forecasting.

Furthermore, the study examines the consistency of performances across different levels of demand lumpiness. It is found that the superior performances of the two-stage models persist when demand shifts from low to moderate levels of lumpiness. However, only the dynamic adaptive and Kalman filter smoothing models retain their good performances at highly lumpy demand level. Other two-stage models involving static exponential smoothing (fixed constants) do not perform up to parity when demand is volatile. The implication is that the forecasting system can handle a certain degree of demand changes without explicit human intervention and that computational intelligence may help alleviate the issue of high demand uncertainty and lumpiness.

NOTES

1. The production facility was closed in the last week of December and the first week of January every year in observance of the holidays.

2. The first year (1997) in the estimation period is reserved as an initialization period for various exponential smoothing models.

REFERENCES

Anderson, D. R., Sweeney, D. J., & Williams, T. A. (2005). *An introduction to management science: Quantitative approaches to decision making* (11th ed.). Mason, OH: Thomson South-Western.

Chen, A. S., & Leung, M. T. (2004). Regression neural network for error correction in foreign exchange forecasting and trading. *Computers and Operations Research*, *31*, 1049–1068.

Fair, R., & Shiller, R. (1990). Comparing information in forecasts from econometric models. *American Economic Review*, *80*, 375–389.

Holt, C. C. (1957). Forecasting seasonal and trends by exponential weighted moving averages. *Office of Naval Research*, Memorandum no. 52.

Mabert, V. A. (1978). Forecast modification based upon residual analysis: A case study of check volume estimation. *Decision Sciences*, *9*, 285–296.

Makridakis, S., & Wheelwright, S. C. (1977). *Interactive forecasting*. Palo Alto, CA: Scientific Press.

Makridakis, S., Wheelwright, S. C., & McGee, V. E. (1983). *Forecasting: Methods and applications* (2nd ed.). New York: Wiley.

Quintana, R., & Leung, M. T. (2007). Adaptive exponential smoothing versus conventional approaches for lumpy demand forecasting: Case of production planning for a manufacturing line. *International Journal of Production Research*, *45*, 4937–4957.

Specht, D. (1991). A general regression neural network. *IEEE Transactions on Neural Networks*, *2*, 568–576.

Theil, H. (1966). *Applied economic forecasting*. Amsterdam: North Holland.

Wasserman, P. D. (1993). *Advanced methods in neural computing*. New York: Van Nostrand Reinhold.

Winter, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management Science*, *6*, 324–342.