Multi-Period Ahead Forecasting with Residual Extrapolation and Information Sharing – Leveraging Multitude of Retail Series

Ozden Gur Ali[1] and Efe Pinar[2]

[1]Department of Business Administration, Koc University, Turkey – oali@ku.edu.tr

[2]Department of Industrial Engineering, Koc University, Turkey

Abstract

Multi-period sales forecasts are important inputs to operations at retail chains with hundreds of stores, many formats, customer segments and categories. Beyond seasonality, holidays and marketing, correlated random disturbances affect sales across stores that share common characteristics.

We propose a novel method, *2 Stage Information Sharing*, that leverages this challenging complexity: Segment-specific panel regressions with seasonality and marketing variables pool the data for better parameter estimates. The residuals are extrapolated non-parametrically using features that are constructed from the last twelve months of observations from the focal and related category-store time series. The final forecast combines the extrapolated residuals with the first stage forecasts.

Working with the extensive dataset of the leading Turkish retailer, we show that the method significantly outperforms panel regression models (mixed model) with AR (1) error structure and the Autoregressive Distributed Lags (ADL) model as well as the univariate exponential smoothing (Winter's) forecasts. The farther out the prediction, the more the improvement.

Keywords

Multivariate time series, Sales forecasting, Panel data, Data mining, Regression, Retail, Multi-period ahead forecast

1. Introduction

Retail forecasts are essential inputs to business decisions in marketing, sales, production, procurement, finance, accounting and human resource management (Mentzer & Bienstock, 1998). Short term (hourly, daily, weekly) demand forecasts at the stock keeping unit (SKU) level drive the procurement and inventory decisions, while long term (multi-year) forecasts of store or chain revenue are essential inputs for capital investment decisions.

In this paper we focus on medium term (up to a year), multi-period (monthly) retail store sales forecasts that constitute critical inputs to the budgeting, resource allocation and incentive compensation calculation processes. Retailers typically have multiple stores of different formats, serving different customer segments in different locations, or even different channels (brick and mortar, internet, mobile). The budgeting and resource allocation process requires objective sales forecasts at the store level and higher, and the capability to evaluate the impact of marketing scenarios. Some of the factors that affect retail sales are within the retail managers' control, such as pricing and promotions; and measuring their impact is critical to efficient resource allocation. Some factors are not controllable but their timing is known – such as seasons and holidays; and understanding their impact allows the managers to design strategies to react in a favorable way. There are many other drivers of retail sales, such as the local and national economy, acts of competition or customer opinion/ sentiment about the company or products, which manifest themselves as random disturbances to sales time series correlated across category-stores that share particular characteristics.

A large portion of the aggregate retail sales forecasting literature deals with univariate time series based on trend, seasonality and autocorrelation structure, e.g. (Alon, Qi, & Sadowski, 2001) and (Chu & Zhang, 2003). Causal models are capable of incorporating the effect of important drivers – such as marketing mix, but reliable estimation of the response parameters in addition to seasonality is challenging, and raises the data availability problem particularly for newer stores/ categories. Even when long time series are available relevance of older data to current dynamics is questioned (Mcintyre, Achabal, & Miller, 1993).

Pooling addresses the data availability issue by leveraging analogous sales time series to learn common patterns e.g.(Bunn & Vassilopoulos, 1999), (Frees & Miller, 2004), and (Lu & Wang, 2010). Pooling observations across stores and subcategories instead of constructing item-store specific models improves the accuracy of regression forecasting models significantly (GürAli, Sayin, van Woensel, & Fransoo, 2009). Econometric models of panel data, which consist of pooled analogous time series,

typically focus on estimating the impact of the drivers efficiently by accounting for the temporal and cross-sectional error structure.

In this paper we propose a two stage approach to multi-period forecasting of multivariate retail sales with covariates that leverages the abundance of data and the business taxonomy for better predictive accuracy: *2 Stage Information Sharing.* The first stage pools series by store segment to estimate the seasonality, calendar and marketing effects in a regression analysis to exploit the high sample size for better parameter estimates. The residuals of this model contain components peculiar to the category-store components that are common to particular groups, such as customer segments or formats, as well as noise. The second stage consists of lead-time specific models that extrapolate the residual time series without assuming a specific error structure, using features constructed with its own recent values, as well as features extracted from the average residual series of groups that are exposed to similar external effects. This approach facilitates information sharing among stores in the second stage models. The idea is that the average residual of the relevant group will be a more efficient estimator of the uncontrolled factors affecting the group, while cancelling irregular effects (noise). The initial forecast is calculated with the Stage 1 regression based on the marketing plan and adjusted using the second stage models for the desired lead time.

The proposed approach differs from existing panel data forecasting methods in the following ways: a) It considers features of a substantial history (12) of random disturbances from all series (stores) that are relevant in some dimension to the focal series (category-store), rather than relying on the estimation algorithm to select the appropriate combination of lags from appropriate stores (series), or requiring the analyst to hand pick them. b) The two stage model fitting with OLS and backward selection is amenable to processing high volumes of series, complex relationships among series and unbalanced panels. c) It uses lead time specific models. The importance of this contribution increases with the size and complexity of the panel data structure. The method allows the analyst to guide the model estimation process by conveying the domain knowledge in terms of features.

We evaluate the proposed forecasting method with the largest retailer of Turkey, with an extensive dataset entailing 363 stores and seven product categories, at the category-store and store levels with 1 to 12 months forecasting lead time. The forecasts constitute the sales expectations of retailer and are used to calculate the incentive component of the store manager compensation objectively, effectively assuming that the deviation from the forecasted sales is due to management effort and practices. Further, the store level forecasts are rolled up for budgeting purposes, and potential drivers of the

deviations from the aggregated forecasts are considered for strategic insights. The proposed method significantly improves the predictive accuracy compared with a Mixed Model with AR(1) error structure and lead-time specific Autoregressive Distributed Lag (ADL) models that use the same inputs and pooling segments as the proposed method; as well as the univariate exponential smoothing (Winter's) forecasts. The improvement in the absolute percentage error compared to the AR (1) Mixed Model is 1.6% for a representative store forecast, and 1.1% for a representative category-store forecast (which correspond to 16% and 8% imrovement in terms of percentage improvement respectively), across lead times. The improvement increases with the forecast lead time as the AR(1) only relies on the last residual of the focal series which is most relevant for immediate forecasts, but ignores the rest of the focal and similar residual series which can provide additional information. Further, the proposed method employs lead-time specific models that allow weighing information differently according to the forecast lead time. The ADL model, which uses the same lags as the proposed model with lead time specific models, has a comparable performance as the proposed model in terms of the *median* absolute percentage error, however 15% of the forecasts have very high (>100%) errors, blowing up the MAPE values at all lead times.

We further show that the added computational complexity due to Stage 2, i.e., extrapolation of the residuals from the panel regression (Stage 1) is justified as it significantly improves the predictive accuracy over Stage 1. Similarly, Information Sharing - across stores within category and across categories within store - significantly improves performance over using only the focal residual series. Finally, including the marketing variables and using store-specific seasonality terms both significantly improve the accuracy of the forecasts.

The rest of the paper is organized as follows. In the next section we review the relevant literature streams. Section 3 describes the retail data characteristics assumed for this work, while section 4 specifies the proposed method. In section 5 we describe the case study and provide accuracy evaluation results compared to external and internal method benchmarks. Section 6 concludes with a summary of contributions, limitations and future research directions. The Appendix contains descriptive statistics for the data, and an illustration of the proposed forecasting method for few category-stores and marketing scenarios.

2. Relevant literature

Aggregate retail sales forecasting deals with sales due to many items, as opposed to item (SKU) level forecasts. Sales can be aggregated in a geographic, product, customer segment, store type, or time

hierarchy. Store, category or category-store level forecasts are aggregate time series because they consist of the total sales (in value) of many items. They can also be aggregated up further to chain level, regional or national forecasts, depending on the organizational needs for the specific situation (Zotteri & Kalchschmidt, 2007).

The category-store level sales also constitute *analogous time series* that can be leveraged for improved accuracy. Analogous time series are subject to similar external factors, such as the economy, competition, retail chain policies or general purchasing patterns (Bunn & Vassilopoulos, 1999). Estimation of seasonality constants using analogous time series improves forecasting accuracy compared to individual time series analysis (Bunn & Vassilopoulos, 1999) and (Chen & Boylan, 2008). Time series can be grouped based on different criteria: based on the business hierarchy, cluster analysis of estimated parameters and cluster analysis of time series (Bunn & Vassilopoulos, 1999).

Duncan et al summarize the benefits of pooling as improved forecasting accuracy with short and noisy time series, fewer parameters to be estimated, adapting rapidly to changes in time series and robustness in the presence of outlier observations (Duncan et al., 2001). They pool analogous time series – which "follow similar time series patterns since they are subject to same or similar consumer tastes, local economic cycles, weather, and regional trends", scale them for the magnitude, construct models considering series trend and level both at the individual and the pooled aggregate level, and combine them such that the final trend and level estimates for the individual series are shrunk toward the aggregate estimate with weights that are inversely proportional to the variance of the estimates. Corberan-Vallet et al argue that series subject to correlated random disturbances do not necessarily have a common structure, and propose a MCMC simulation procedure for the exponential smoothing model of multiple series (Corberán-Vallet, Bermúdez, & Vercher, 2011).

The econometrics literature defines multiple *pooled* related time series as panel data, and uses regression analysis with the main objective of consistent and efficient estimation of the impact of various factors on the response variable, for example for policy analysis, (Greene, 2008; Wooldridge, 2009). The major model types are the pooled regression with a common intercept and variable parameter for all time series, the fixed effects model with a group specific intercept, the random effects model with group specific random element, and the random parameters model that (provided there is enough data) allows representation of the heterogeneity in the variable parameters. The marketing literature uses econometric models of panel data to infer promotion response with choice models e.g.(Erdem, 1996), and (Guadagni & Little, 1983). There is a vast literature in marketing focusing on the

correct estimation of the promotion and price response, exploring different aspects e.g., endogeneity of marketing decisions (Chintagunta, Dubé, & Singh, 2003), dynamics of marketing and consumer decisions (Dekimpe & Hanssens, 2000) and (Pauwels, 2004), asymmetric effects of price thresholds (Pauwels, Srinivasan, & Franses, 2007), store level elasticities (Hoch, Kim, Montgomery, & Rossi, 1995), and cross-category elasticities (Kamakura & Kang, 2007).

A criticism of econometric panel regression methods is that they are designed for controlling for the "nuisance" variation while estimating causal models (Duncan et al., 2001). This criticism appears to be in line with the arguments that modeling and forecasting are distinct activities (Allen & Fildes, 2001), and that the causal models often predict less accurately than naïve formulations due to model misspecification interacting with irregularities in the economy (Chevillon & Hendry, 2005).

Frees and Miller (2004) observe that the panel data regression models are hardly used for forecasting purposes, and show how trend slopes that vary by subject or time, serial correlation, or random walk can be represented in a longitudinal data mixed model. They forecast lottery sales by location up to five weeks ahead with pooled cross-sectional, error components, fixed effects and two way error component models with serially uncorrelated and correlated, i.e., AR(1) error structures (Frees & Miller, 2004). Specifically in retail forecasting (beverage sales by channels and regions), Divakar et al use random effects models allowing the price and temperature impact to vary among regions/channels, while keeping the errors i.i.d. (Divakar, Ratchford, & Shankar, 2005).

GurAli et al (2009) experiment with an extensive multi-store, multi-category 76 week long grocery store dataset to identify the impact of pooling for one-step ahead forecasting of SKU-store sales with marketing mix variables. Their findings indicate that pooling across stores and subcategories improves the forecasting accuracy significantly regardless of the regression method: stepwise multiple regression, regression tree, and support vector regression. They further propose a regression tree approach with the pooled data using static and dynamic descriptors of the SKU, store and category characteristics, which results in 65% improvement in the forecasting accuracy of the time periods with promotions (GürAli et al., 2009). An L1-norm regularized epsilon insensitive regression with similar data provides simpler models (Gür Ali, 2013). Huang et al., also point to the problem of too many explanatory variables and use Lasso (which constrains the L1 norm of coefficients) or factor analysis with an ADL model with lags of sales and explanatory variables to forecast SKU level sales with promotion and competitive information (Huang, Fildes, & Soopramanien, 2014). Fildes et al., compare the forecasting accuracy of several econometric models and find that pooled ADL models perform better than the

vector autoregressive (VAR), time varying parameter (TVP) models and the univariate ADL, AR(3) and exponential smoothing models (Fildes, Wei, & Ismail, 2011).

Another approach to leveraging related time series for improved forecasting is to cluster time series according to some measure and then construct machine learning models for each identified group of time series. Lu and Wang (2010) cluster demand time series from the computer industry from the same time period, and build a support vector regression (SVR) model for each group. The mixing matrix that specifies the weights of the independent components in each demand time series is used to cluster the demand time series into disjoint clusters using Growing Hierarchical Self Organizing Maps. Each cluster is fit a separate SVR with optimized parameters. When forecasting, a classification algorithm is used to identify the appropriate cluster before using the appropriate SVR model (Lu & Wang, 2010).

Scalability of the forecasting method is important when dealing with real retail applications with hundreds of stores, thousands of SKUs. One aspect of scalability is the amount of analyst time required to build and maintain the forecasting model: this becomes a serious concern when correct specification for hundreds of econometric models is desired (Huang, Fildes, & Soopramanien, 2014). Another aspect of scalability is whether the method can be estimated with large datasets. For example, application of the celebrated support vector regression method on a pooled dataset with dynamic and static descriptors of the SKU, store, and category proved to be problematic due to the size of the dataset and memory limitations. The ROCSA (Row and Column Selection Algorithm) enables SVR model estimation by selectively subsampling the data to keep important observations and variables and which also improves forecasting accuracy for noisy datasets (Gür Ali & Yaman, 2013).

Multi-step ahead forecasting is an important requirement for planning in the retail industry; the forecasts are updated as new information becomes available. On the other hand, the vast majority of the forecasting literature is concerned with one-step-ahead forecasts. In multi-step forecasting there are two main approaches: iterated use of the one-step ahead forecasting (IMS), and direct multi-step estimation (DMS) where the model is specifically estimated to minimize the specific multi-step error. Chevillon and Hendry provide a review of the considerable literature debating when each method is more appropriate (Chevillon & Hendry, 2005), and point out that misspecification of the error process (e.g. the degree of the AR or MA process) results in DMS being more accurate. Based on simulation experiments they conclude that the non-parametric DMS (where each lead time is a different model, rather than one model with a lead time parameter for multiple horizons) results in potential gains, and is

more robust compared to the IMS approach, particularly in the presence of varying trends and cyclical patterns.

3. Retail Chain Data Characteristics

Retail chains are characterized by a multitude (hundreds) of stores of different formats in diverse geographic regions. Stores of the same format have similar size, layout and product assortment. Stores serve different customer segments of socio-economic status by virtue of their specific location. Customer segments are not nested within the format or region. Rather, each region and most formats contain stores serving several customer segments.

A category consists of many SKUs and brands. Category-store sales are the aggregated value of all SKUs in the category at the given store. Higher level forecasts are obtained by aggregating sales forecasts of the relevant stores. Most, but not all stores carry all categories. The length of the historical time series is store specific. Hence the panel data is unbalanced.

The problem we address here is forecasting the store and higher (such as format or region) level sales up to L time periods into the future, given the planned levels of marketing variables at the category-store level.

There are hundreds of stores, thousands of category-store combinations, and a multitude of forecasting horizons for each time series, rendering a manual approach to model selection and testing impractical. On the other hand, the large volume of data makes it possible to identify and leverage the local trends in specific segments that may be too subtle in individual time series. Beyond seasonality, calendar and marketing effects, sales are affected by random disturbances. The source of these random disturbances may be national, thus affecting all stores and categories; format-specific, customer segment specific, region specific, or store specific. Examples are as follows: *national and companywide issues:* changes in national consumer confidence or shopping habits; entry, exit or growth of national competition, changes in the incentive compensation; *format specific issues*: changes in the layout, assortment, item availability, or specific competition that are carried out in all stores of the format; *customer segment specific issues*, such as changes in the consumer confidence, shopping habits or competition specific to this segment; *category specific issues*, such as changes in the assortment, customer tastes, or category specific competition e.g. due to specific internet sites; *store specific issues*, such as changes in the demographics due to migration or urban transformation, new traffic patterns/ construction in the local area, change in store personnel.

The large number of category-store sales time series offers a valuable information source for detecting the format, customer segment, category, region or store specific signals, which can then be used in forecasting the disturbances.

4. The Method

The proposed *2 Stage Information Sharing* mehod consists of two stages: the first stage expresses the store specific category sales as a function of the seasonality and calendar effects and marketing variables, while the second stage models extrapolate the residuals from the stage one model multiple steps ahead by using the recent residuals of the focal category-store, as well as residual series of category-stores that share common characteristics with the focal category-store.

Since the marketing spend levels and the customer responses differ by category as well as store characteristics, we forecast at the category-store level, rather than the store level, and aggregate up to store and higher levels.

The first stage equation provides the initial sales predictions given the marketing plans and the calendar. These rough estimates are adjusted using the second stage models. Figure 1 provides a graphical overview of our forecasting approach.

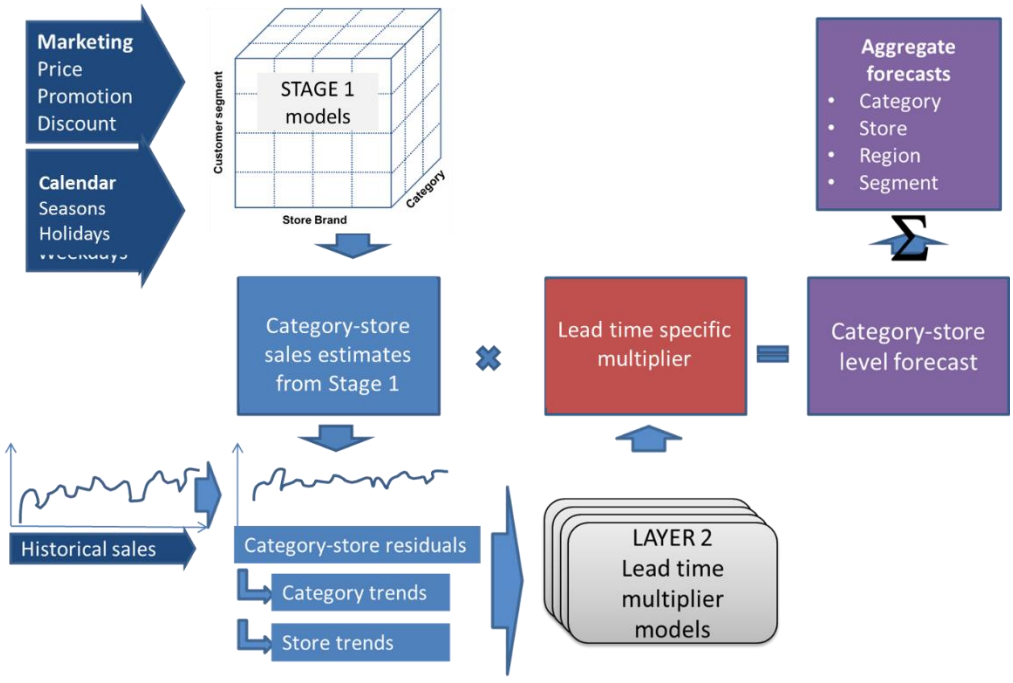

Figure 1. Overview of the forecasting approach

4.1. First stage model adjusting for marketing and calendar effects

$$\ln y_t = \alpha + \boldsymbol{\beta H_t} + \boldsymbol{\gamma M_t} + \varepsilon_t$$

The first stage models the $\log_e$, i.e.; ln of store category-specific sales, expressed in value, $\ln y_t$, at time t as a linear model of the seasonality and calendar variables, the vector $\boldsymbol{H}_t$, and the marketing variables $\boldsymbol{M}_t$. The marketing variables may include the lags, as many marketing papers conclude that marketing actions have an effect in the future periods (Hanssens, Parsons, & Schultz, 2003). We use the ln of marketing variables and ln of sales as established in the marketing literature e.g. (Hoch et al., 1995). ε_t is the residual term, which is likely to be serially correlated.

This is a straight forward log-linear formulation of the category-store level sales. However, the number of parameters, compared with the number of observations makes it difficult to estimate the model at the individual category-store level. Each year of data provides only one data point for the estimation of seasonality parameters, while the marketing variables may not vary too much across time, large peaks occurring seldom. Individually fitting such a model with OLS regression, frequently results in unrealistic parameter estimates, as discussed in e.g. (Bemmaor, Franses, & Kippers, 1999)

Therefore, we pool category sales time series of similar stores to increase the number of observations, as shown in equation (1). A key issue in leveraging analogous time series for forecasting is the tradeoff between bias and variance. For example, pooling normalized sales data across stores and estimating a model that assumes a common trend parameter will have a variance decreasing effect due to increased sample size, and an aggregation bias arising due to differences within the pooled series. Empirical approaches to determining the level of pooling include correlational co-movement group or clustering locations based on their characteristics, and expert judgment (Duncan et al., 2001). We pool the stores according to the drivers of promotion response, since the model assigns the same set of marketing response parameters to all stores in the pool. Consistent with the marketing literature, we expect the sales response to marketing to differ based on the customer affluence, the assortment and the product category e.g. (Bijmolt, Van Heerde, & Pieters, 2005) and (Pauwels et al., 2007), thus we estimate a model for each category, store format and customer segment. Seasonality is observed to differ by store even within the geographical region, hence store-specific seasonality parameters are introduced within the pooled model.

$$\ln y_{ijt} = c_{ij} + \boldsymbol{\beta_{ij} H_t} + \boldsymbol{\gamma_j M_{ijt}} + \varepsilon_{ijt} \qquad (1)$$

In equation (1) y_{ijt} stands for the sales of store *i*, category *j* at time *t*. The fixed effect, c_{ij}, adjusts for the level differences of individual category-store sales. The category specific row vector $\boldsymbol{\gamma_j}$, contains the category marketing effect parameters common to the pooled stores. The heterogeneity of the store seasonality and calendar effects among the pooled stores is accounted for by the category-store specific parameter vector $\boldsymbol{\beta_{ij}}$.

This is a longitudinal data mixed model (Frees & Miller, 2004) with fixed effects, where some covariate (the seasonality and calendar variable) parameters are cross-section specific, while other covariate (marketing variables) parameters are common. The model does not have time specific terms, as we would like the time specific effects to be explained by the covariates, and those that are not explained by the covariates to be reflected in the residuals, which will be further analyzed by the stage two models. It is not necessarily balanced; i.e., not all cross-sections have the same time series length, since some stores have longer history than others.

As long as the independent variables are exogenous[1], even though the errors ε_{ijt} are not assumed to be independent, the parameters can be estimated consistently, although not efficiently, with the OLS approach (Greene, 2008). The large sample size due to pooling to estimate the parameters should help improve the efficiency of the estimators.

4.2. Second stage lead time specific models

The residuals of the first stage model ε_{ijt}, constitute sales time series for store *i* and category *j* that are adjusted for seasonality, holidays, and marketing effects. We extrapolate the residual time series with lead time and segment specific pooled regression models. From a statistical learning/ machine learning point of view, we use variables/ features that summarize the information in the residual time series of the focal series as well as similar series based on domain knowledge: similarity of stores in different dimensions and time series components.

Compared to the AR(1) error term structure which relies only on the last observed residual for the focal time series to forecast the residuals, this model uses more information by considering a) the residuals up to a year old, and b) the residuals from similar time series. Further, the lead-time specific models

[1] A potential source of endogeneity is that retailers may determine the level of marketing variables (e.g., set prices) in response to or in expectation of sales levels, which will cause the estimated impact of the marketing variable to be biased(Leeflang & Wittink, 2000). Approaches to deal with the endogeneity problem include designed experiments where the retailer deliberately sets different prices in similar locations under similar situations, and the instrumental variables which must be correlated with the endogenous explanatory variable, while not being correlated with the error terms.

allow weighing information differently according to the forecast lead time. For example, we would expect that the trend observed in the past couple of months to be relevant for the next two month's forecast, however when forecasting a year ahead it would make more sense to consider the trend over the past year. While more information and lead-time specific model estimation allow for a better fit with a more flexible model, they also open the door for overfitting[2] the specific patterns observed in the single time series. To counter this possibility we pool the residual time series for the estimation of the regression coefficients. We use the same segmentation as in Stage 1, which is based on product category, customer segment and store format.

Extrapolation of a time series relies on estimates of the time series' current level, trend estimates and adjustment for recent deviation from seasonality. Even though we have adjusted for seasonality in the first stage model, the residual from last year may represent a new development that can manifest itself again in the same season this year. We use the last two observations for level, and the last observation of the month to be predicted for seasonality. We use estimates of the recent change in the level of the residuals at different leads.

$$Delta_k(\varepsilon_t) = \frac{\sum_{m=0}^{k-1} \varepsilon_{t-m}}{k} - \frac{\sum_{m=0}^{k-1} \varepsilon_{t-k-m}}{k}$$

This feature provides the change in the average level of the time series in the last k time periods compared with the previous k time periods.

Own residuals only

$$\varepsilon_{ij,t_0+l} = \theta_{l0} + \theta_{l1}\varepsilon_{ijt_0} + \theta_{l2}\varepsilon_{ij\,t_0-1} + \theta_{l3}Delta_2(\varepsilon_{ijt_0}) + \theta_{l4}Delta_3(\varepsilon_{ijt_0}) + \theta_{l5}Delta(\varepsilon_{ijt_0}) + \theta_{l6}\varepsilon_{ij,+l-12} + v_{ij,t_0+l} \quad (2)$$

Equation (2) displays the simple version of the second stage model that models the category-store specific residuals of equation (1) at time t_0+*l*, using the information that is available as of t_0. It uses the following features of the residual time series: the last two observed residuals, the last observed residual of the predicted month, and estimates of the change in the level of the residuals with the last four, six

[2] A model that overfits describes the noise in the data rather than the underlying relationship, increasing the in sample fit while reducing the hold-out accuracy. This typically occurs when complex models are applied to data of insufficient size and variability (Hastie, Tibshirani, & Friedman, 2009).

and twelve observations[3]. Notice that in equation (2) when *l*=11 $\varepsilon_{ij,t_0+l-12} = \varepsilon_{ij,t_0-1}$ and when *l*=12 $\varepsilon_{ij,t_0+l-12} = \varepsilon_{ijt_0}$, and the term with the last observed residual of the predicted month falls out.

Equation (3) provides a more compact representation of the same equation using a feature set definition as follows.

$$\varepsilon_{ij,t_0+l} = \theta_{l0} + \boldsymbol{\Theta_l}\, Features(\boldsymbol{\varepsilon_{ij}}) + v_{ij,t_0+l} \qquad (3)$$

Here, $\boldsymbol{\Theta_l}$ is a row vector of parameters and $Features(\boldsymbol{x})$ is a column vector with the following elements:

$Feature_1(\boldsymbol{x}) = x_{t_0}$; $Feature_2(\boldsymbol{x}) = x_{t_0-1}$; $Feature_3(\boldsymbol{x}) = Delta_2(x_{t_0})$;

$Feature_4(\boldsymbol{x}) = Delta_3(x_{t_0})$; $Feature_5(\boldsymbol{x}) = Delta_6(x_{t_0})$; $Feature_6(\boldsymbol{x}) = x_{t_0+l-12}$;

where $\boldsymbol{x}$ consists of the 12 most recent time series observations x_{t_0-11} to x_{t_0}.

Notice that a given category-store has a different function to extrapolate the residual for each lead time *l*. Even though the available input variables for each model are the same, the actual variables selected for the model and their coefficients are different by lead time. The irregular component of the category-store time series is denoted by v_{ijt} .

Information sharing

Average residuals across stores with a particular characteristic, are expected to reflect the impact of the common factors affecting those stores at that time period, while reducing the variability due to store specific issues. Thus, in addition to category-store specific residual time series, the average residual time series of stores that share a common characteristic with the focal store, or the average residual time series of the focal store across all categories can also be considered for extrapolation. We use the same six features to summarize these additional time series.

4.3. Fitting the models

For both the stage 1 and 2 models we use weighted least squares regression that favors the more recent observations over the older ones. The weight of the observation at time *t*, w_t, is calculated similar to the weights given to observations in exponential smoothing, as follows.

$$w_t = \alpha * (1-\alpha)^{T-t}$$

[3] One could introduce features with other k values as well. In the interest of reducing the potential for overfitting, we left out k=4 and 5.

α is a constant between 0 and 1, the higher its value the more weight given to more recent observations, and *T* is the length of the time series.

The number of models that need to be constructed makes it necessary to automate the forecasting procedure. Further, the number of store-specific seasonality terms in Stage 1 and the number of potential variables in Stage 2 call for elimination of unnecessary variables to guard against overfitting. While more sophisticated feature selection techniques can be employed, considering practical implementability we use the commonly used backward elimination procedure to drop the insignificant variables. In the stage 1 model, the store effects (c_{ij}) and the common coefficients of the marketing variables ($\boldsymbol{\gamma_j}$) are required, while the store-specific seasonality effects ($\boldsymbol{\beta_{ij}}$) are only kept in the model if they are significant. Hence, in Stage 1 the backward elimination is only applied to the store-specific seasonality. In the Stage 2 models, our null hypothesis is that there is no particular error structure in the data, hence all terms are subject to backward elimination.

4.4. Forecasting

The forecast for store *i*, category *j* sales at t+l are calculated as follows

$$\hat{y}_{ij,t_0+l} = \exp(\hat{\alpha}_{ij} + \widehat{\boldsymbol{\beta}}_{\boldsymbol{ij}}\boldsymbol{H}_{t_0+\boldsymbol{l}} + \widehat{\boldsymbol{\gamma}}_{\boldsymbol{j}}\,\boldsymbol{M}_{\boldsymbol{ij}\,t_0+\boldsymbol{l}} + \hat{\varepsilon}_{ij,t_0+l}) \qquad (4)$$

Here $\boldsymbol{H}_{t_0+l}$ represents the holiday and seasonal variable values at t time t_0+*l*, which are known in advance, and $\boldsymbol{M}_{ij,t_0+l}$ contains the planned level of marketing variables for store *i* category *j* at time t_0+*l*, set by the managers. $\widehat{\boldsymbol{\alpha}}_{ij}, \widehat{\boldsymbol{\beta}}_{ij}, \widehat{\boldsymbol{\gamma}}_{j}$ are the parameters estimated by the stage 1 model. $\hat{\boldsymbol{\varepsilon}}_{ij,t_0+l}$ is the residual estimate for time period t_0+*l*, calculated with the stage 2 model , using the residual time series of the focal category-store *ij*, and the average residual time series of stores similar in particular dimension.

Our experiments with adjusting for the log transformation bias using (1) the parametric (Miller, 1984) or (2) nonparametric methods (Duan, 1983) did not provide significant improvement to the holdout forecast accuracy, hence we opted for the non-adjusted, lower complexity forecasting equation.

5. Application to retail chain

We apply the method to forecast store level monthly sales one to twelve months ahead for the largest retailer of Turkey. The retailer would like to use the forecasts at the store and higher levels of aggregation, such as at the region, format, and national level for the budgeting process, and to provide

the critical input for manager incentive compensation calculations. They would like to be able to adjust the price, discount and customer specific promotion levels and estimate the impact on sales.

5.1. The data

The stores across the country are organized under four store formats. Formats differ in terms of their store concept and assortment; two formats are further divided into three subformats according to the size of the store. Stores can also be segmented according to the socio-economic status of the clientele that they serve in the particular location into A, B and C groups. Two of the formats serve all three clientele segments, whereas one format serves high-end customers exclusively. There are six geographic regions. There are seven product categories, including fresh foods with considerable seasonality, packaged foods, meat products, milk products, cleaning products and cosmetics, and the sublet section. The data comprises 60 months of 336 stores with seven categories that have been open for more than two years, starting with January 2007. In total there are 2330 category-store time series, since some stores may not carry all categories. Their lengths range from 37 to 60 months and have an average of 58.2 months and standard deviation of 5.3 months. Hence, the data is not balanced in terms of length or categories. Tests do not indicate nonstationarity[4]. The sales are expressed in value (TL)[5]. For confidentiality purposes, the sales amounts have been multiplied by a constant. The basic statistics and correlations are provided in the Appendix I. To ensure high sample size and thus reduce the impact of particular events we use the rolling origin approach to accuracy evaluation. There are five origins (August – December 2010) where the test data consists of the following twelve months for each origin. We pool the stores according to the store subformat, customer segment and product category, yielding 133 Stage 1 and Stage 2 models. We estimate the Stage 1 models as specified in (1), where the vector $\boldsymbol{H}_t$ consists of 11 dummy variables for the months and 4 variables for the number of different types of holidays and special days in the month. The number of data points available to estimate an element of the category-store specific parameter vector $\boldsymbol{\beta}_{ij}$ for a particular month is limited to the number of years of data for the store. On the other hand, the observed seasonality of the stores exhibits significant variation from store to store, even within the same geographic region – as can be seen from the example in Appendix II, and our experimentation with pooled seasonality parameters resulted in significantly worse holdout accuracy than store specific parameters as can be seen in Table 3 . Notice

[4] We ran the Lm-Pasaran-Shin unit root test for the sales time series,, which tests the hypothesis that all panels contain unit roots against the hypothesis that some panels are stationary. The null hypothesis was rejected at <0.00001 level.

[5] Although marketing models typically use sales in units, the aggregate number of units sold in a category is not useful for practical purposes, as the items vary greatly in value.

that the alternative method of estimating the seasonality coefficients as a pretreatment step would have ignored seasonality of the marketing variables and potentially underestimated the impact of marketing.

The vector $\boldsymbol{M}_{ijt}$ contains the price index, the level of discounts, and the level of CRM (Customer Relationship Management) promotions for store *i*, category *j* and time period *t*. The price level is an index that reflects the average price of the items in the category assortment of the store format at the time. It is not adjusted for the store's price level versus its competitors. The prices of individual items are the same across stores in the same format, since they are determined by the headquarters, but they vary in time. The discount and CRM promotion levels are indices of the average *planned* temporary price reduction and the average *planned* CRM promotion levels, respectively. The CRM offers are available only to targeted customers, unlike the discounts that are accessible to all. The discount and CRM levels for category *j* at time *t* are the same across stores in the same format and geographical region.[6]

As explained earlier, the marketing variables can affect the sales not just in the time period they are applied, but also in the subsequent periods. We have evaluated versions of the proposed Stage 1 model with and without the lags of the marketing variables, and found that including the lags of the marketing variables did not change the holdout accuracy significantly[7]. Hence, we opted for the simpler model without the lags of the marketing variables.

In the Stage 2 models we evaluate forecasts with own information as specified in (3) as well as with information sharing, specified as follows.

$$\varepsilon_{ij,t_0+l} = \theta_{l0} + \boldsymbol{\Theta_l}\, Features(\boldsymbol{\varepsilon_{ij}}) + \boldsymbol{\Omega_l}\, Features(\boldsymbol{\varepsilon_{\bar{d}j}}) + \boldsymbol{\Xi_l}\, Features(\boldsymbol{\varepsilon_{i.}}) + \boldsymbol{\Phi_l}\, Features(\boldsymbol{\varepsilon_{.j}}) + v_{ij,t_0+l} \quad (5)$$

Here $\varepsilon_{\bar{d}jt}$ is the average category *j* residual at time *t* across stores that share characteristic *d* with the focal store, in this case - those that are in the same format, $\varepsilon_{i.t}$ is the average residual at time *t* for store *i* across categories, and $\varepsilon_{.jt}$ is the average category *j* residual at time *t* across all stores. All input time

[6] Since our models do not deal with the SKU level data, we use indices at the format, region and category level, where the SKU discounts are set uniformly across stores, and the demand is expected to be similar. The indices can be calculated as the average planned discount for the SKU (and the targeted customer group for CRM) weighted by the historical SKU share. There are many combinations of SKU discounts and campaigns levels that can result in the same index value, and translating the index to particular SKU discounts and campaigns is beyond the scope of this work.

[7] We have used five rolling origins as explained in section 5, and found that the MAPE in the holdout data was 0.01% higher with the lags of the marketing variables, however this is not a significant difference (p value=0.36).

series are described with the same set of features as in (3): the last two observations at t_0 and t_0-*1*, the last observation from the predicted season t_0+*l*-*12,* and trend estimates based on last 4, 6 and 12 months.

In this work, keeping with the exponential smoothing literature we have used the common value 0.05 for α, which implies an observation that is two years old weighs about 30% of a current observation. The backward elimination threshold is set at 0.10.

We provide an illustration of the method with few series in Appendix II.

5.3. Stage 2 model parameters

The Stage 2 models have a large number of potential terms, but about half are eliminated by the backward elimination procedure. Figure 2 and Figure 3 provide the average percent of the Stage 2 models containing the group of terms by lead time, averaged over the five evaluation origins. Figure 2 groups the terms according to the data source; we observe that the average residual series at the category and category-format levels are used slightly more frequently than the focal category-store residual series, which in turn are selected more frequently than the store level residuals averaged over categories. One explanation for the lower usage percentage of the store terms is that the information content regarding the category is more useful than the store specific content; another explanation is that the store level averages are noisier since they contain an average over 7 categories, while category-format or category level averages average over tens to hundreds of stores. As expected, as the lead time increases, fewer terms are found to make a sufficiently significant contribution to extrapolating the category-store level residual series. Interestingly, the inclusion rate of the less noisy category and category-format level residual series declines much slower than the focal residual series inclusion rate. In other words, information sharing terms constitute a relatively higher percentage of the terms in the longer (than shorter) lead time forecasting models.

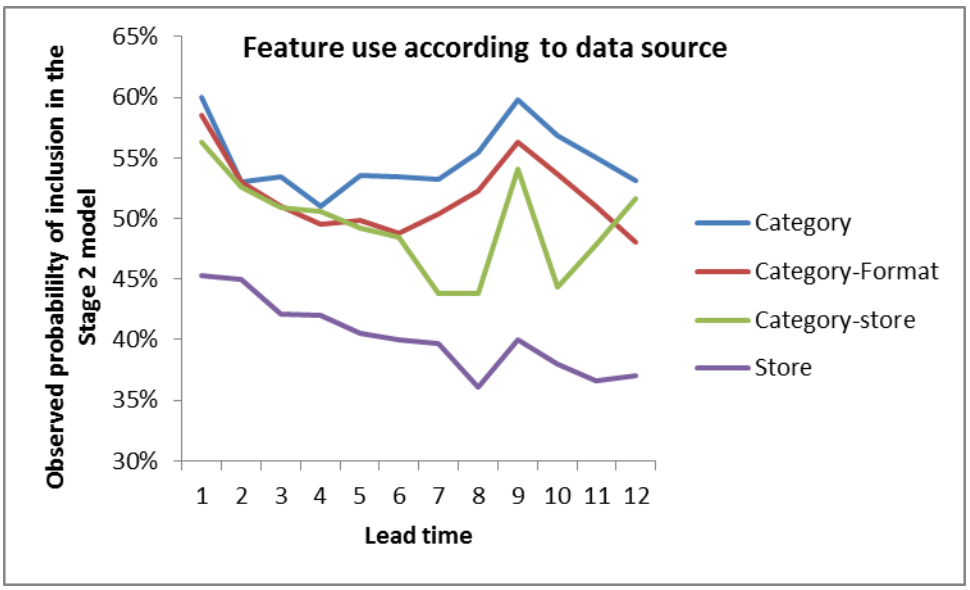

Figure 2. Residual series usage in Stage 2 models by lead time

Figure 3 groups the terms according to the feature. For example, we observe that while residuals from the last observed time period, t_0, are selected most frequently for forecasting one to three months ahead, the six-monthly Delta feature turns out to be the most useful for longer lead times. We observe that the last two periods' residuals (t_0 and t_0-*1*) become less important as the lead time increases. The last residual from the predicted season (t_0*+l-12*) is selected on average in about half the models, although we have already accounted for the systematic seasonality in Stage 1 model. This last residual from the predicted season provides information about a new development that may manifest itself again in the same season this year; in other words, a change in the established seasonality. The 2 on 2 and 3 on 3 month Delta terms are less popular than the other features, but still are present in about 40 to 45% of the models.

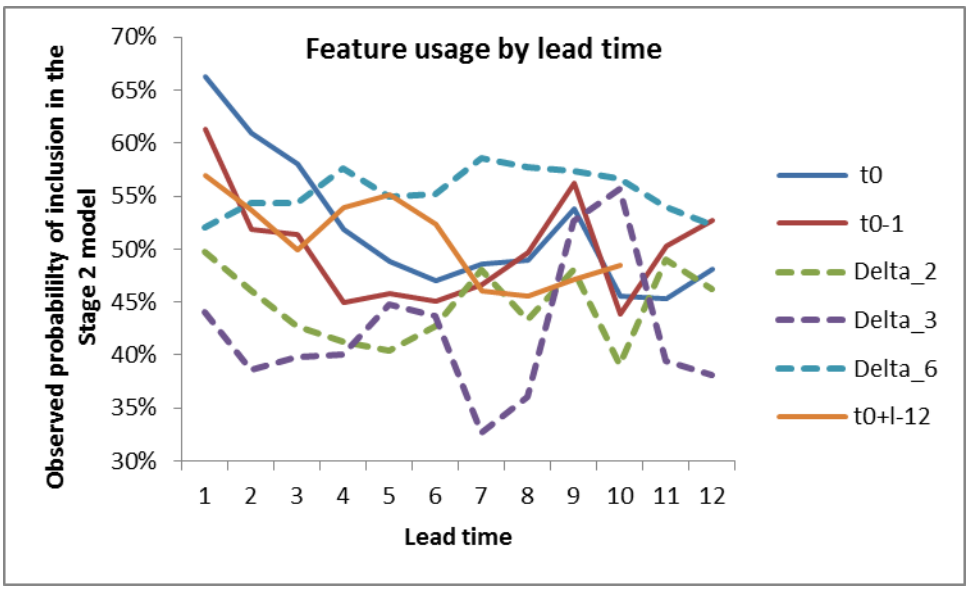

Figure 3. Feature usage in Stage 2 models by lead time

5.4. Accuracy evaluation

In this section we compare the forecasting accuracy of the proposed model with the benchmark methods by lead time and aggregation level. We further conduct experiments to identify the contribution of method components and information sources.

Compared methods

We use three external benchmarks to compare the accuracy of forecasts by the proposed method:

1) The Winter's method is the representative of simple univariate models that incorporate seasonality. It is a member of the exponential smoothing family that is well established in practice and shown to perform well in diverse forecasting situations.

2) A mixed model with AR(1) error component e.g. see Frees and Miller (2004), that uses the same pooling groups and variables as the proposed method Stage 1 model is specified similar to the Stage 1 model as follows.

$$ln\, y_{ijt} = c_{ij} + \boldsymbol{\beta_{ij} H_t} + \boldsymbol{\gamma_j M_{ijt}} + \varepsilon_{ijt}$$

$$\varepsilon_{ijt} = \rho\, \varepsilon_{ij,t-1} + \delta_{ijt}$$

Here $|\rho| < 1$, and δ_{ijt} is i.i.d., with mean 0 and variance σ_δ^2. The forecasts are calculated as follows.

$$ln\, \hat{y}_{ij,t_0+l} = c_{ij} + \boldsymbol{\beta_{ij} H_t} + \boldsymbol{\gamma_j M_{ijt}} + \rho^l \varepsilon_{ij,t_0} \qquad (6)$$

Therefore, the forecasts of the mixed model with AR(1) error component model rely only on the *last* observed residual of the *focal* time series to incorporate the random error.

3) Lead time specific ADL models with the same explanatory variables as the proposed Stage 1 model and twelve lags of the dependent variable. The model is specified as follows, and fitted using the same the same pooling groups as the proposed model, using backward elimination of terms and OLS.

$$\ln y_{ij\,,t_0+l} = c_{ij} + \boldsymbol{\beta_{ij}} \boldsymbol{H}_{t_0+l} + \boldsymbol{\gamma_j}\, \boldsymbol{M_{ij},}_{t_0+l} + \sum_{k=0}^{11} \boldsymbol{\Psi} \ln y_{ij\,,t_0-k} + \varepsilon_{ij\,,t_0+l} \tag{7}$$

Further, to evaluate the incremental accuracy due to method components, we compare the accuracy of the forecasts made with the following models: a) Stage 1, b) 2 Stage with own residual series, and c) 2 Stage Information Sharing.

Finally, to identify the accuracy impact of the marketing variables and store specific seasonality we compare the accuracy of a) the Stage 1 model with vs without marketing variables, b) the Stage 1 model with store specific versus pooled seasonality, and c) 2 Stage Information Sharing with store specific versus pooled seasonality models.

Accuracy measures

We report the accuracy of all methods at the category-store level, at which they are estimated, as well as at the store level at which they are used. The store level forecasts are calculated by aggregating the category-store sales forecasts.

As accuracy measures, we report the mean absolute percentage error (MAPE), the mean absolute error (MAE) and the median absolute percentage error (MdAPE). MAPE is a popular measure that summarizes the relative error across all observations, while MAE is relevant because it captures the actual magnitude of the forecasting error. Even though two categories of a store may have the same MAPE, the MAE values can be substantially different due to category size. MAE weights observations based on their size, while MAPE weighs them equally. MdAPE is robust to potential outliers, and provides the perspective of the representative store; however it can provide an unduly positive picture if observations with very large errors are not just occasional outliers, but they are systematically present.

Results

Tables 1, 2 and 3 provide the average differences in terms of all accuracy measures between methods by lead time, aggregation level and overall. The summaries that are provided in the figures and tables are across series, not just within series observations. Each store level data point is based on a sample size of

336 stores and 5 origins per month; providing 5040 observations in lead time groups 1-3 and 4-6 months and 10080 observations for the 7-12 month lead time group. The sample size of the category-store level measures is approximately seven times higher, as each store carries at most seven categories.

Overall accuracy results

Figure 4 and Figure 5 provide the overall holdout accuracy of the methods averaged across lead times, in terms MAPE, MdAPE and MAE at the store and category-store levels, respectively. While the MAE values at the store level are higher due to the size effect, the relative errors are uniformly lower than the category-store level forecasts for all methods and lead times. This is in line with the expectations, assuming that the models are unbiased and the errors from individual components are independent from each other.

The first observation based on Figure 4 and Figure 5 is that the proposed method (2 Stage Information Sharing) has better overall accuracy than the benchmarks, Winter's exponential smoothing and the mixed model with AR(1) error structure, in terms of all three accuracy measures for store level and category-store level forecasts. The rows marked "Overall" in Table 1 indicate that all differences are statistically significant. For a representative store, the improvement in the absolute percentage error due to using the proposed method versus the mixed model is 16%, and for a representative category-store the improvement is 8%[8].

The MAPE and MAE values of the other benchmark model, ADL, "blow up", while its MdAPE values are similar to the proposed method. Trying to explain this phenomenon we questioned whether ADL had few outlier observations that skewed the results and produced the observed MAPE and MAE values in the order of 10^{13} to 10^{19}. We found that the ADL model produced high error rates[9] for a substantial, i.e.; 15% of the stores and category- stores, rendering it not useful for forecasting purposes. The twelve lags in the ADL model, in addition to the marketing and seasonality variables create an environment that fosters overfitting and results in a very wide range of coefficient estimates for the marketing variables reaching implausibly high values, such as 25 for the ln of price index – which then produce the exploding forecasts. In contrast, the first stage of the proposed 2 Stage Information Sharing method, or the AR(1) mixed model, have 12 (11) fewer parameters involved in the estimation process of the marketing variables. Another problem with the leadtime specific ADL models is that they produce twelve

[8] The overall MdAPE for the proposed model is 13.0% at the category-store level and 8.7% at the store level; for the mixed model the figures are 14.2% and 10.3% respectively.

[9] Arbitrarily defined as absolute percentage error > 100%

independent estimates of the marketing impact for the same category and store segment, which are frequently substantially different.

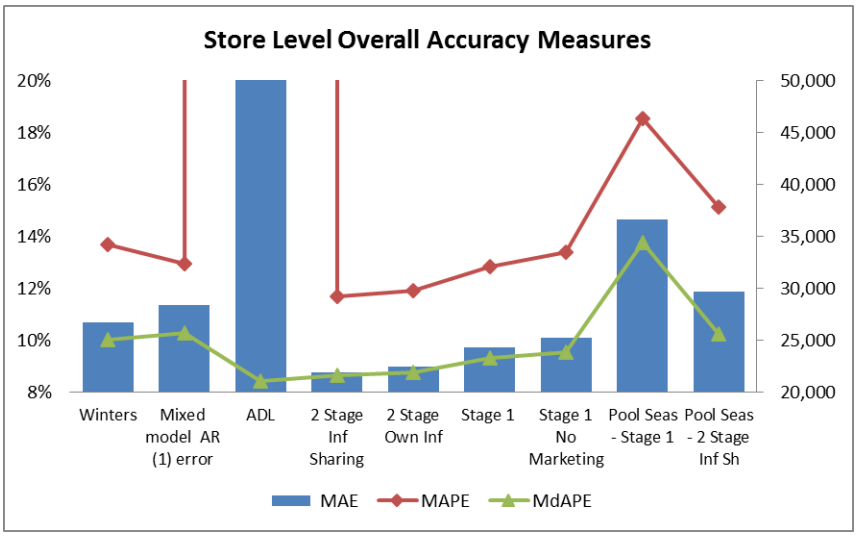

Figure 4. Overall accuracy comparison of all methods across all lead times, at the store level. Each figure corresponds to an average over 12 lead times, 5 rolling horizons and 336 stores. The ADL average MAPE and MAE values are too high to fit to the graph: 6.3 10^{13} and 3.9 10^{19} respectively.

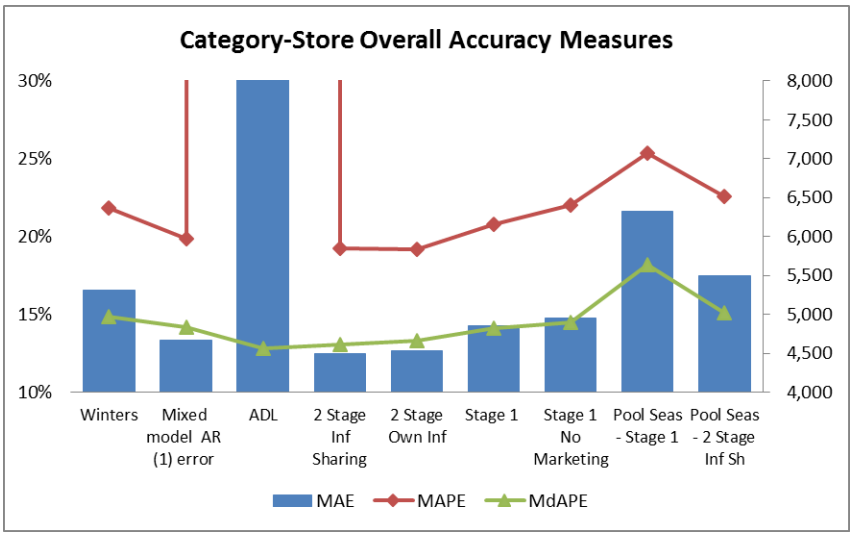

Figure 5 Overall accuracy comparison of all methods across all lead times, at the category-store level. Each figure corresponds to an average accuracy measures over 12 lead times, 5 rolling horizons, 336 stores and seven categories. The ADL average MAPE and MAE values are too high to fit to the graph: 2.1 10^{14} and 5.6 10^{18}

To explore why the accuracy difference between the proposed method and the mixed model AR(1) is greater at the store level than the category-store level forecasts, we take a look at the category-store level accuracy of the models by sales size buckets (Figure 6). As expected, the MAPE decreases with

category-store size for both methods. Interestingly, the proposed method provides a substantial accuracy improvement over the benchmark Mixed model for all but the smallest category-stores[10]. This can potentially be explained with the signal to noise ratio of the residual time series increasing with the size of the category-store, thus improving the accuracy of the proposed method. Figure 7 shows that these larger category-stores contribute 85% of sales, but account for only 45% of the category-stores. Hence, the proposed model shows a larger improvement over the Mixed model for the store level forecasts than the category-store level forecasts.

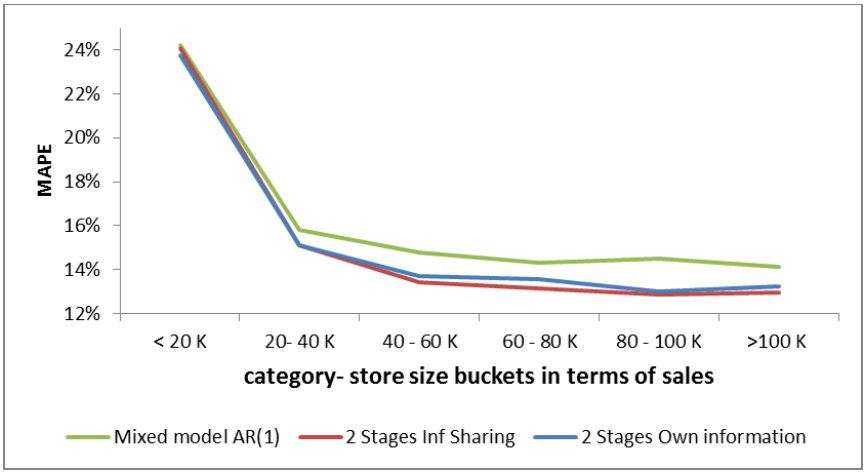

Figure 6. Category-store level MAPE of the proposed 2 Stages Information Sharing, 2 Stages Own Information and the benchmark Mixed Model AR (1) methods by the category-store size buckets.

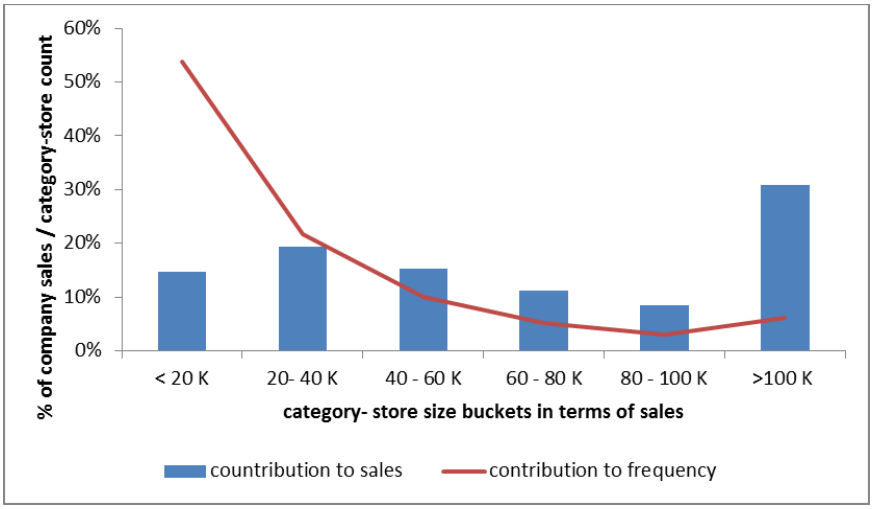

Figure 7. The contribution of the category-store buckets to company sales and their frequency as a percent of the company sales and company category-store count, respectively

[10] Note that the actual sales figures have been scaled to preserve data confidentiality.

Accuracy results by lead time

Figure 8 and Figure 9 report the holdout MAPE and MdAPE respectively for the category-store and store level forecasts by lead time[11]. ADL is not included in these figures since the MAPE and MAE values are too high to fit the graph. As expected, we observe that as the lead time increases, all error measures increase for each method. The proposed 2 Stage Information Sharing model performs significantly better than the Winter's model for all lead times, as can be seen in Table 1. Interestingly, as the lead time increases the accuracy difference between the proposed 2 Stage Information Sharing model and the Mixed model increases in favor of the Information Sharing model. For a representative store, the accuracies are not significantly different for the 1-3 months lead time window; however for longer lead times the 2 Stage Information Sharing model is significantly better, providing 11% lower error for 4-6 months lead time and 18% lower error for the 7-12 months lead time horizon[12]. For a representative category-store, the Mixed model is better in the 1-3 months lead time horizon, while the Information Sharing model provides more improvement in the 7-12 months lead time horizon.

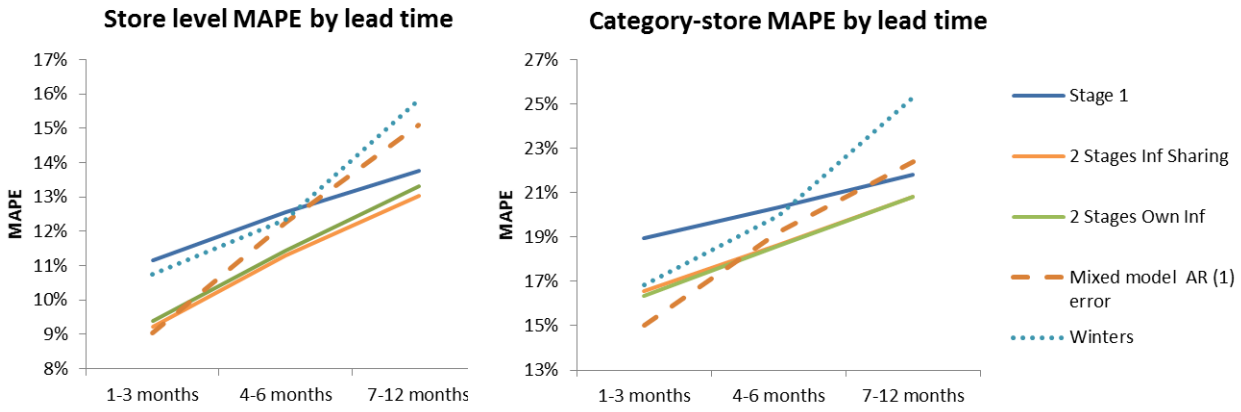

Figure 8. Average MAPE of store and category-store level forecasts by lead time bucket for the proposed and benchmark methods.

[11] MAE results are similar and hence are left out in the interest of saving space.

[12] The store level MdAPE for the proposed model is 8.3% for 4-6 months lead time, and 9.0% for 7-12 months lead time; for the Mixed model the figures are 9.3% and 10.9% respectively.

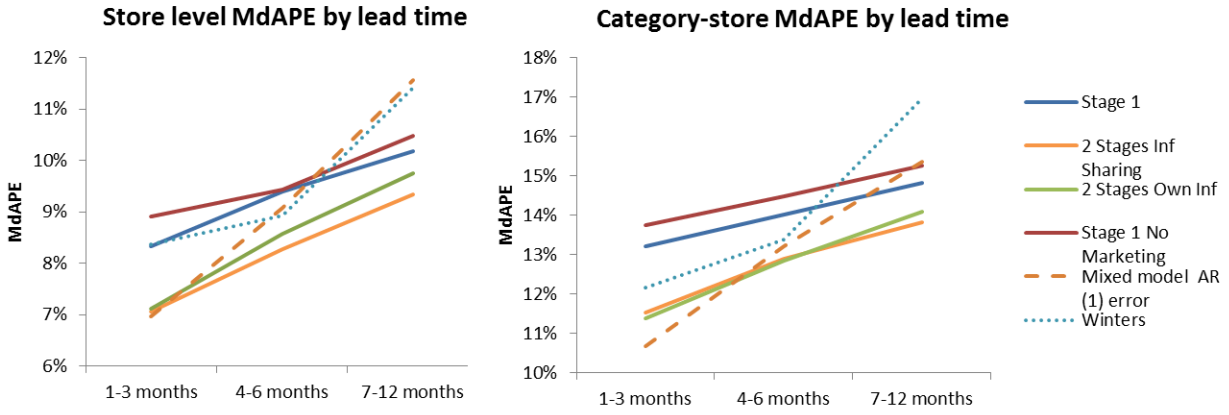

Figure 9. Average MdAPE of store and category-store level forecasts by lead time bucket for the proposed and benchmark methods.

Table 1. Accuracy comparison of the proposed method with external benchmarks. Average differences between methods are provided in terms of MAPE, MdAPE and MAE by leadtime and overall, for category-store and store level forecasts. All the category-store level differences are significant at the most at the 0.05 level, unless in italics; and all the store level differences are significant at the most at the 0.01 level, unless in italics . The differences with the ADL model have p-values in the 0.16 range due to the very high variability of the ADL accuracy.

Compared methods		Winter's exp smoothing – 2 Stages Inf Sharing			Mixed model AR(1) - 2 Stages Inf Sharing			ADL model - 2 Stages Inf Sharing		
Aggregation	**Lead time**	**MAPE**	**MdAPE**	**MAE**	**MAPE**	**MdAPE**	**MAE**	**MAPE**	**MdAPE**	**MAE**
None (Category-store level)	**1-3 months**	*0.0%*	0.4%	196	-1.4%	-1.2%	-211	1.07E+02	-1.0%	1.22E+07
	4-6 months	1.3%	0.4%	402	0.6%	*0.0%*	*35*	8.60E+14	-0.6%	2.22E+19
	7-12months	4.5%	3.2%	1315	1.6%	1.2%	429	5.35E+09	0.2%	5.88E+14
	Overall	2.6%	1.8%	817	0.6%	1.1%	177	2.19E+14	-0.3%	5.65E+18
Store level	**1-3 months**	1.3%	1.1%	3150	*-0.1%*	*-0.3%*	*3147*	1.38E+02	*0.2%*	7.88E+07
	4-6 months	1.1%	0.6%	2621	1.0%	1.0%	4788	2.53E+14	*-0.1%*	1.54E+20
	7-12months	2.8%	2.1%	6455	2.0%	1.9%	8894	5.98E+09	*0.0%*	4.07E+15
	Overall	2.0%	1.5%	4696	1.3%	1.6%	6486	6.34E+13	*0.0%*	3.85E+19

Contribution of stages

Investigating the contribution of the Stage 2 beyond the Stage 1 model in the proposed method, we see in Table 2 that both versions of Stage 2, i.e., Information Sharing using equation (5) and Own Information using equation (2) significantly improve the predictive accuracy over Stage 1 overall and for all lead times, at the store and category-store levels. We observe from Figure 8, Figure 9 and Table 2 that the accuracy improvement arising from the Stage 2 model over the Stage 1 model decreases as expected with the lead time, as new disturbances during the lead time diminish the relevance of the extrapolated historical residuals.

The Information Sharing version has significantly better overall predictive accuracy than the Own Information version of the proposed method, based on Table 2. On the other hand, the 2 Stage with Own Information version performs significantly better than the 2 Stage Information Sharing at the category-store level for short lead times (1-3 months), while the Information Sharing version does better for the 7-12 months lead time horizon, and for all lead times at the store level. In general, we observe that the contribution from Information Sharing improves accuracy beyond the Own Information version particularly for longer lead times and store level forecasts.

Table 2. Internal accuracy comparison of the proposed method. Average differences between methods are provided in terms of MAPE, MdAPE and MAE by leadtime and overall, for category-store and store level forecasts. All the category-store level differences are significant at the most at the 0.05 level, unless in italics; and all the store level differences are significant at the most at the 0.01 level, unless in italics .

Compared methods		**Stage 1 - 2 Stages Inf Sharing**			**2 Stages Own Inf - 2 Stages Inf Sharing**			**Stage 1 - 2 Stages Own Inf**		
Aggregation	**Lead time**	**MAPE**	**MdAPE**	**MAE**	**MAPE**	**MdAPE**	**MAE**	**MAPE**	**MdAPE**	**MAE**
None (Category-store level)	**1-3 months**	2.3%	1.0%	464	-0.2%	-0.4%	*-12*	2.6%	1.5%	476
	4-6 months	1.7%	1.4%	318	*0.0%*	*0.2%*	*-19*	1.7%	1.2%	337
	7-12 months	1.0%	1.1%	317	*0.0%*	0.3%	91	1.0%	0.9%	226
	Overall	1.5%	1.1%	352	*-0.1%*	0.2%	38	1.6%	0.9%	314
Store level	**1-3 months**	1.9%	1.2%	3677	0.2%	*0.1%*	471	1.7%	1.1%	3205
	4-6 months	1.3%	1.1%	2366	*0.2%*	0.5%	*240*	1.1%	0.6%	2126
	7-12 months	0.7%	0.9%	1707	0.3%	0.4%	754	0.4%	0.5%	952
	Overall	1.1%	0.6%	2342	0.2%	0.1%	556	0.9%	0.5%	1785

Accuracy impact of marketing and store-specific seasonality variables

As can be seen in Table 3, the inclusion of the marketing variables significantly improves the accuracy of the Stage 1 model predictions for all measures and lead times, with one exception. For a representative store, the impact accounts for about 3% improvement in the accuracy at both the store and category-store levels. This may be an underestimation of the marketing impact as the marketing variables are to a large extent seasonal (see Figure 11), and therefore their impact may be partially accounted for by the seasonality terms in the Stage 1 model. Store specific seasonality terms improve the model accuracy significantly compared to pooled seasonality for both Stage 1 and 2 models, for all lead times (see Table 3).

Table 3 . Accuracy impact of the marketing variables and store specific seasonality. Average differences between methods are provided in terms of MAPE, MdAPE and MAE by leadtime and overall, for category-store and store level forecasts. All the category-store level differences are significant at the most at the 0.05 level, unless in italics; and all the store level differences are significant at the most at the 0.01 level, unless in italics

		Impact of Marketing variables			Impact of store specific seasonality					
					Stage 1			2 Stages Inf Sharing		
Compared methods		Stage 1 No Marketing - Stage 1			Pooled seasonality - Store specific seasonality			Pooled seasonality - Store specific seasonality		
Aggregation	Lead time	MAPE	MdAPE	MAE	MAPE	MdAPE	MAE	MAPE	MdAPE	MAE
None (Category-store level)	1-3 months	1.1%	0.9%	53	3.82%	4.87%	678	3.10%	3.6%	400
	4-6 months	1.3%	0.6%	26	3.85%	4.05%	875	2.71%	2.5%	426
	7-12 months	1.3%	0.0%	166	5.49%	3.65%	2194	3.89%	2.0%	1603
	Overall	1.3%	0.4%	104	4.68%	3.96%	1499	3.40%	2.4%	1019
Store level	1-3 months	0.5%	0.5%	586	6.11%	7.10%	8626	3.97%	4.27%	5115
	4-6 months	0.2%	*0.5%*	350	5.48%	5.82%	8647	3.08%	2.60%	4220
	7-12 months	0.8%	*0.4%*	1457	6.07%	4.87%	16903	3.73%	2.38%	11636
	Overall	0.6%	0.2%	969	5.93%	5.50%	12770	3.63%	2.57%	8152

6. Conclusions, limitations and future research

We proposed and evaluated a multi-period forecasting method, *2 Stage Information Sharing*, that leverages multiple analogous (retail sales) time series to improve the forecast accuracy. The forecast consists of two components: the expected sales that capture the category-store specific seasonality, marketing plans and the segment specific marketing response; and the extrapolated residuals, sharing information about recent random disturbances in similar locations.

The method uses the concepts of pooling time series for parameter estimation, time series decomposition (into calendar, marketing and trend-cyclical components) and direct multi-step estimation in a unique design that differentiates it from other approaches. However, the main contributions are the non-parametric, feature driven extrapolation of the recent residuals with lead time specific models, and information sharing among stores using the average disturbances in relevant groups of category-stores.

The extrapolation is non-parametric in the sense that no particular error structure is assumed, unlike e.g., the AR (1) model that assumes serially correlated errors. In the proposed method, the extrapolation model can include terms from the features constructed with the most recent year's residuals, i.e., the last two observations, three local trend estimates with different horizons, and the last observation of the relevant month, as the complexity of the observed data requires. Providing these relevant features reduces the search space and the chances of overfitting, compared with using the raw residual time series as input.

We evaluated the accuracy of the category-store and (aggregated) store level forecasts with one to twelve months lead time using data from the leading Turkish retailer; specifically, 336 stores with seven categories in four formats. The proposed method outperforms the benchmark models, the Mixed model with an AR(1) error structure, the lead-time specific Autoregressive Distributed Lags (ADL) models – both having the same explanatory variables and pooling groups as the proposed model, as well as the univariate exponential smoothing (Winter's) forecasts . The accuracy improvement compared to the Mixed model and Winter's method are higher for longer lead times. The ADL model results in MAPE values in the order of 10^{13} to 10^{14} due to unacceptably high errors for 15% of the observations, along with highly variable estimates for marketing variables that are inconsistent across lead times.

We show that the second stage model that extrapolates the residuals of the focal series significantly improves forecast accuracy beyond the first stage model with calendar and marketing effects. Adding

Information Sharing in the second stage model improves the accuracy further beyond using only the focal category-store residual time series, particularly for longer lead times and store level forecasts. A limitation of the evaluation is that we have applied the method to a specific retailer's data, although the dataset contains a large number of stores from different formats, serving diverse customer segments in various geographies. The first stage models are subject to potential endogeneity of the marketing decisions, as the retailer may determine the level based on recent sales or anticipated events such as competitor's actions.

The Stage 1 and Stage 2 models can be potentially improved by employing different regression formulations, or even resorting to more flexible techniques such as neural nets or support vector regression. However, this would require addressing the transparency and interpretability concerns of such techniques.

Further research can explore the feature and data source selection problem in the second stage models with machine learning methods, such as a regularized regression that penalizes complexity as well as errors, or a wrapper approach that adds data sources and/ or features based on validation data set accuracy.

Finally, another interesting idea, suggested by an anonymous reviewer, is to use the judgmental adjustments to forecasts (see e.g., Trapero, Pedregal, Fildes, & Kourentzes, 2013) in addition to the residuals as inputs to the extrapolation of random disturbances, as they contain adjustments due to anticipated events as well.

Appendix I

Basic statistics and correlations between sales and marketing variables

Variable	Mean	Std Dev
$\ln(Y_{ijt})$	9.64	1.35
$\ln(CRM_{ijt})$	-6.02	1.27
$\ln(Discounts_{ijt})$	-5.70	1.50
$\ln(PriceIndex_{ijt})$	0.07	0.09

	$\log_e(Y_{ijt})$	$\log_e(CRM_{ijt})$	$\log_e(Discounts_{ijt})$
$\ln(CRM_{ijt})$	0.09		
$\ln(Discounts_{ijt})$	0.06	0.73	
$\ln(PriceIndex_{ijt})$	0.08	0.12	0.18

Appendix II

Illustration of the method

In this section we illustrate the method with few time series. Figure 10 provides time series plots of four category-store sales of the same category, format and customer segment. Stores 1 and 2 are in Region 1 and stores 3 and 4 are in Region 2. Notice that the seasonality patterns can be quite different, even for stores in the same geographic region. Therefore we use store-specific, rather than region-specific seasonality terms. The relevant time series of the marketing variables are provided in Figure 11; we see that there is considerable temporal variation in the marketing variables, but the variation across regions is minimal. Here, observations 49-60 constitute the test data, while time periods 1-48 are used for estimating the models.

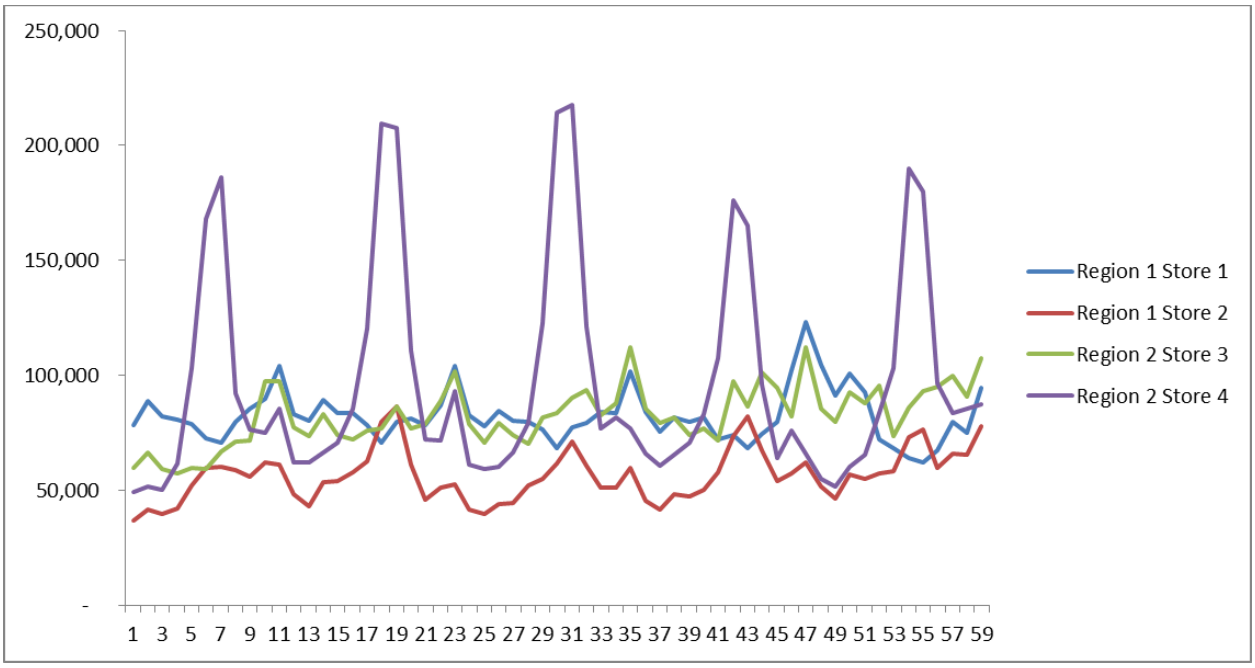

Figure 10. Selected category-store sales time series

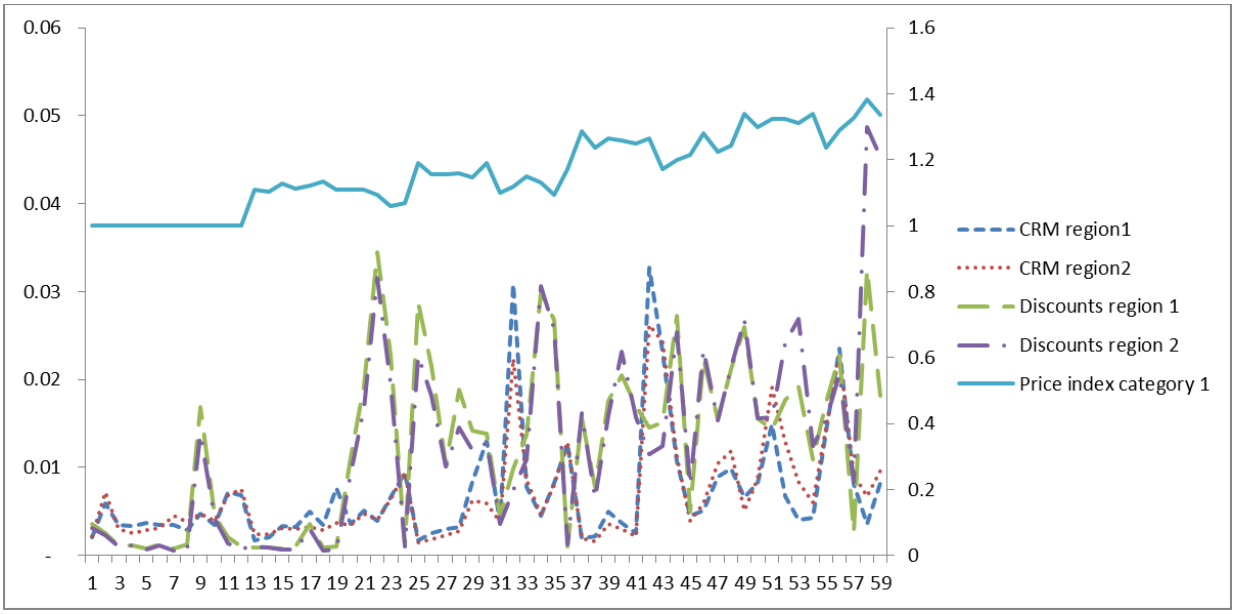

Figure 11. Time series of the marketing variables for the selected category-stores

Table 4. Stage 1 model relevant coefficients for the selected category-store series

	Common	Region 1 Store 1	Region 1 Store 2	Region 2 Store 3	Region 2 Store 4
Store effect		11.39	10.91	11.47	11.22
feb			-0.17	-0.13	
apr				-0.16	
may				-0.16	0.20
jun			0.16	-0.14	0.54
jul		-0.12	0.33		1.03
aug			0.46		1.03
sep			0.25		0.44
nov		0.16	0.11		
dec		0.35	0.19	0.23	
holiday_1					1.22
$\log_e$_price_index	0.20				
$\log_e$ crm	0.02				
$\log_e$ discount	0.01				

The coefficients of the Stage 1 model for this segment are provided in Table 4. The coefficients for the marketing variables are significant at the 0.005 level, while the seasonality effects were selected by backward elimination with a 0.1 max significance level requirement. The model has an RMSE of 0.012. Interpreting the seasonality parameters, we observe e.g., that Store 4 has higher sales in summer and holidays, while Store 1 sales are higher in winter. The positive price parameter indicates that as the price increases the sales in this segment increase in *value;* but the parameter is less than 1, hence the increase in value is due to higher prices making up for the loss in *unit sales*. Particularly, considering that the average yearly consumer price index inflation rate during the study period was 8.7% [13], the consumers - to a certain extent – expect nominal price increases[14]. We also observe that sales in this segment are more responsive to the customer specific CRM promotions than the discounts. Using these coefficients, the residuals for the training time period are calculated.

Next, the average residual series across categories for each store, and the average residual series across stores for the relevant category and format-category are calculated (as illustrated in Figure 12). We observe that the average residuals at the category and format-category level are smoother and oscillate within smaller bands than the category-store level series. These time series are used to calculate the input features for the twelve second stage lead time specific regression models, as described by equation (5), which uses the average residual series across the stores for the specific category, category-format, and across the categories in the specific store, in addition to the category-store residual series.

[13] Source: Turkish Statistical Institute http://www.turkstat.gov.tr/

[14] Our experimentations with inflation adjusted prices did not result in better models.

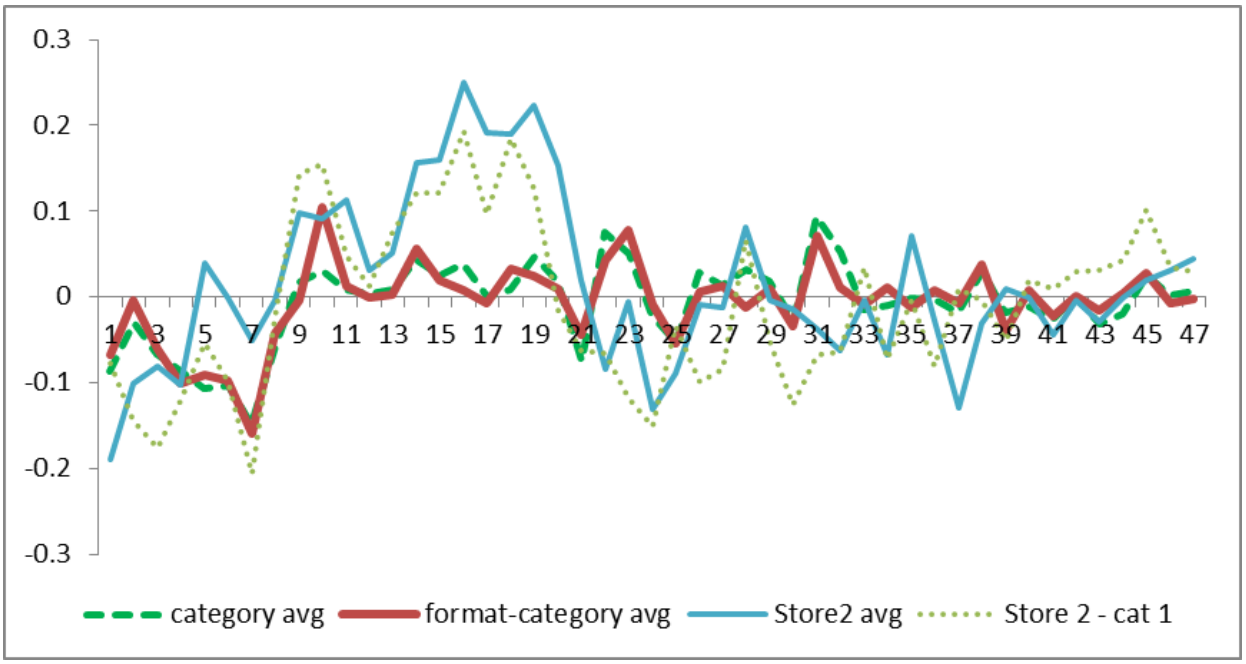

Figure 12. Category-store specific residuals for Store 2 category 1, and the averaged residuals across categories for the store, across stores for the category and format-category.

Figure 13 provides the category-store level predictions for the holdout time period in all categories of Store 2, assuming the base case marketing plans provided in Figure 11. Figure 13 also illustrates an additional marketing scenario that entails 3 percentage point additional CRM and discount for each month in the planning horizon for category 1, which increases the sales forecasts versus the base case marketing scenario in category 1 by 3 to 6% according to the month.

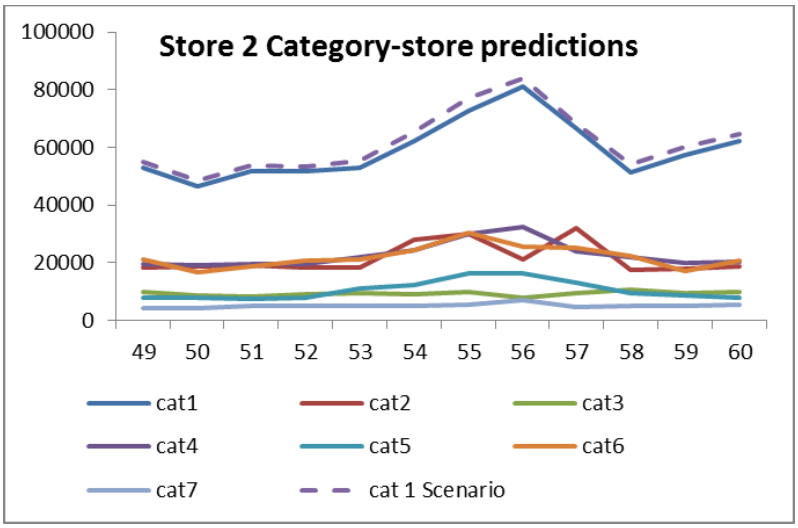

Figure 13. Store 2 category level forecasts. Category 1 Scenario entails additional 3 percentage points in CRM and discounting for category1.

References

Allen, P. G., & Fildes, R. (2001). Econometric forecasting In J. S. Armstrong (Ed.), *Principles of Forecasting: A Handbook for Researchers and Practitioners* (pp. 303-362). Boston: Kluwer

Alon, I., Qi, M., & Sadowski, R. J. (2001). Forecasting aggregate retail sales:: a comparison of artificial neural networks and traditional methods. *Journal of Retailing and Consumer Services, 8*(3), 147-156.

Bemmaor, A. C., Franses, P. H., & Kippers, J. (1999). Estimating the impact of displays and other merchandising support on retail brand sales: partial pooling with examples. *Marketing Letters, 10*(1), 87-101.

Bijmolt, T. H. A., Van Heerde, H. J., & Pieters, R. G. M. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of Marketing Research, 42*(2), 141-156. doi: DOI 10.1509/jmkr.42.2.141.62296

Bunn, D. W., & Vassilopoulos, A. I. (1999). Comparison of seasonal estimation methods in multi-item short-term forecasting. *International Journal of Forecasting, 15*(4), 431-443. doi: http://dx.doi.org/10.1016/S0169-2070(99)00005-9

Chen, H., & Boylan, J. E. (2008). Empirical evidence on individual, group and shrinkage seasonal indices. *International Journal of Forecasting, 24*(3), 525-534. doi: http://dx.doi.org/10.1016/j.ijforecast.2008.02.005

Chevillon, G., & Hendry, D. F. (2005). Non-parametric direct multi-step estimation for forecasting economic processes. *International Journal of Forecasting, 21*(2), 201-218. doi: http://dx.doi.org/10.1016/j.ijforecast.2004.08.004

Chintagunta, P. K., Dubé, J.-P., & Singh, V. (2003). Balancing profitability and customer welfare in a supermarket chain. *Quantitative Marketing and Economics, 1*(1), 111-147.

Chu, C.-W., & Zhang, G. P. (2003). A comparative study of linear and nonlinear models for aggregate retail sales forecasting. *International Journal of Production Economics, 86*(3), 217-231. doi: http://dx.doi.org/10.1016/S0925-5273(03)00068-9

Corberán-Vallet, A., Bermúdez, J. D., & Vercher, E. (2011). Forecasting correlated time series with exponential smoothing models. *International Journal of Forecasting, 27*(2), 252-265. doi: http://dx.doi.org/10.1016/j.ijforecast.2010.06.003

Dekimpe, M. G., & Hanssens, D. M. (2000). Time-series models in marketing:: Past, present and future. *International Journal of Research in Marketing, 17*(2), 183-193.

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*(2), 151-167. doi: http://dx.doi.org/10.1016/j.ijpe.2004.02.004

Divakar, S., Ratchford, B. T., & Shankar, V. (2005). Practice Prize Article—CHAN4CAST: A Multichannel, Multiregion Sales Forecasting Model and Decision Support System for Consumer Packaged Goods. *Marketing Science, 24*(3), 334-350.

Duan, N. (1983). Smearing Estimate: A Nonparametric Retransformation Method. *Journal of the American Statistical Association, 78*(383), 605-610. doi: 10.2307/2288126

Duncan, G. T., Gorr, W. L., & Szczypula, J. (2001). Forecasting analogous time series *Principles of forecasting* (pp. 195-214): Springer US.

Erdem, T. (1996). A dynamic analysis of market structure based on panel data. *Marketing Science, 15*(4), 359-378.

Fildes, R., Wei, Y., & Ismail, S. (2011). Evaluating the forecasting performance of econometric models of air passenger traffic flows using multiple error measures. *International Journal of Forecasting, 27*(3), 902-922.

Frees, E. W., & Miller, T. W. (2004). Sales forecasting using longitudinal data models. *International Journal of Forecasting, 20*(1), 99-114. doi: http://dx.doi.org/10.1016/S0169-2070(03)00005-0
Greene, W. H. (2008). *Econometric analysis* (6th ed.). Upper Saddle River, N.J.: Prentice Hall.
Guadagni, P. M., & Little, J. D. (1983). A logit model of brand choice calibrated on scanner data. *Marketing Science, 2*(3), 203-238.
Gür Ali, Ö. (2013). Driver Moderator Method For Retail Sales Prediction. *International Journal of Information Technology & Decision Making, 12*(6), 1261-1286.
Gür Ali, Ö., & Yaman, K. (2013). Selecting rows and columns for training support vector regression models with large retail datasets. *European Journal of Operational Research, 226*(3), 471-480. doi: http://dx.doi.org/10.1016/j.ejor.2012.11.013
GürAli, Ö., Sayin, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*(10), 12340-12348.
Hanssens, D. M., Parsons, L. J., & Schultz, R. L. (2003). *Market response models: Econometric and time series analysis* (Vol. 12): Springer.
Hastie, T., Tibshirani, R., & Friedman, J. H. (2009). *The elements of statistical learning : data mining, inference, and prediction* (2nd ed.). New York, NY: Springer.
Hoch, S. J., Kim, B.-D., Montgomery, A. L., & Rossi, P. E. (1995). Determinants of store-level price elasticity. *Journal of Marketing Research*, 17-29.
Huang, T., Fildes, R., & Soopramanien, D. (2014). The value of competitive information in forecasting FMCG retail product sales and the variable selection problem. *European Journal of Operational Research, 237*(2), 738-748.
Kamakura, W. A., & Kang, W. (2007). Chain-wide and store-level analysis for cross-category management. *Journal of Retailing, 83*(2), 159-170.
Leeflang, P. S., & Wittink, D. R. (2000). Building models for marketing decisions:: Past, present and future. *International Journal of Research in Marketing, 17*(2), 105-126.
Lu, C.-J., & Wang, Y.-W. (2010). Combining independent component analysis and growing hierarchical self-organizing maps with support vector regression in product demand forecasting. *International Journal of Production Economics, 128*(2), 603-613.
Mcintyre, S. H., Achabal, D. D., & Miller, C. M. (1993). Applying Case-Based Reasoning to Forecasting Retail Sales. *Journal of Retailing, 69*(4), 372-398. doi: Doi 10.1016/0022-4359(93)90014-A
Mentzer, J. T., & Bienstock, C. C. (1998). *Sales forecasting management : understanding the techniques, systems, and management of the sales forecasting process*. Thousand Oaks: Sage Publications.
Miller, D. M. (1984). Reducing transformation bias in curve fitting. *The American Statistician, 38*(2), 124-126.
Pauwels, K. (2004). How dynamic consumer response, competitor response, company support, and company inertia shape long-term marketing effectiveness. *Marketing Science, 23*(4), 596-610.
Pauwels, K., Srinivasan, S., & Franses, P. H. (2007). When do price thresholds matter in retail categories? *Marketing Science, 26*(1), 83-100.
Wooldridge, J. M. (2009). *Introductory econometrics : a modern approach* (4th ed.). Mason, OH: South Western, Cengage Learning.
Zotteri, G., & Kalchschmidt, M. (2007). A model for selecting the appropriate level of aggregation in forecasting processes. *International Journal of Production Economics, 108*(1–2), 74-83. doi: http://dx.doi.org/10.1016/j.ijpe.2006.12.030