International Journal of Forecasting 32 (2016) 502–517

Contents lists available at ScienceDirect

International Journal of Forecasting

journal homepage: www.elsevier.com/locate/ijforecast

Multi-period-ahead forecasting with residual extrapolation and information sharing — Utilizing a multitude of retail series

Ozden Gur Ali *, Efe Pinar

Koc University, Istanbul, Turkey

ARTICLE INFO

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

ABSTRACT

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

We propose a novel method, *Two-Stage Information Sharing* that takes advantage of this challenging complexity. In this method, segment-specific panel regressions with seasonality and marketing variables pool the data, in order to provide better parameter estimates. The residuals are then 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 forecasts from the first stage.

Working with the extensive dataset of a leading Turkish retailer, we show that this method significantly outperforms both panel regression models (mixed model) with an AR(1) error structure and the autoregressive distributed lags (ADL) model, as well as the univariate exponential smoothing (Winters') method. The further out the prediction, the greater the improvement.

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 decisions relating to procurement and inventory, while long term (multi-year) forecasts of the 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) sales forecasts from retail stores, which constitute critical inputs for the budgeting, resource allocation and incentive compensation calculation processes. Retailers typically have multiple stores of different formats that serve different customer segments in different locations, and possibly 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 ability to evaluate the impacts of different marketing scenarios. Some of the factors that affect retail sales are within the control of retail managers (such as pricing and promotions), and measurements of their impacts are critical for efficient resource allocation. Other factors are not controllable, but their timing is known (such as seasons and holidays), and an understanding of their impact allows the managers to design favorable strategies in reaction. There are also many other drivers of

* Corresponding author.
E-mail address: oali@ku.edu.tr (O. Gur Ali).

http://dx.doi.org/10.1016/j.ijforecast.2015.03.011
0169-2070/© 2015 International Institute of Forecasters. Published by Elsevier B.V. All rights reserved.

retail sales, such as the local and national economy, acts of competition, and customer opinion/sentiment about the company or products, which manifest themselves as random disturbances to sales time series which are correlated across category-stores that share particular characteristics.

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

Pooling addresses the issue of data availability by using analogous sales time series to determine 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 (Gür Ali, Sayin, van Woensel, & Fransoo, 2009). Econometric models of panel data, which consist of pooled analogous time series, typically focus on estimating the impacts of the drivers efficiently by accounting for the temporal and cross-sectional error structure.

In this paper, we propose a two-stage approach to the multi-period forecasting of multivariate retail sales with covariates that takes advantage of the abundance of data and the business taxonomy in order to enhance the predictive accuracy, which we refer to as *Two-Stage Information Sharing*. The first stage estimates the seasonality, calendar and marketing effects in a regression analysis by pooling the series by store segment, in order to exploit the large sample size and obtain better parameter estimates. The residuals of this model contain components that are peculiar to the category-store and components which 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 any specific error structure, using both features constructed based on its own recent values and 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 that affect the group, while cancelling irregular effects (noise). The initial forecast is calculated via 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 the features from a substantial history (12) of random disturbances from all series (stores) that are relevant to the focal series (category-store) in some way, rather than either 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 the processing of large volumes of series, complex relationships among series, and unbalanced panels.

(c) It uses lead-time-specific models. The importance of this point increases with the size and complexity of the panel data structure. The method allows the analyst to guide the model estimation process by providing their domain knowledge in terms of features.

We evaluate the proposed forecasting method on the largest retailer in Turkey, with an extensive dataset covering 363 stores and seven product categories, at the category-store and store levels, with a forecasting lead time of one to 12 months. The forecasts are adopted by the retailer as sales expectations, and are used to provide an objective calculation of the incentive component of the store manager's compensation, effectively assuming that any deviations from the forecasted sales are due to management efforts and practices. Further, the store-level forecasts are rolled up for budgeting purposes, and potential drivers of any deviations from the aggregated forecasts are examined for strategic insights. The proposed method improves the predictive accuracy significantly compared with either a mixed model with an AR(1) error structure or lead-time-specific autoregressive distributed lag (ADL) models that use the same inputs and pooling segments as the proposed method; as well as with univariate exponential smoothing (Winters') 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 (corresponding to 16% and 8% improvements, respectively, in terms of percentage improvements), across lead times. This improvement increases with the forecast lead time, as the AR(1) relies only on the last residual of the focal series, which is most relevant for immediate forecasts, and 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 information to be weighted 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 performance comparable to that of the proposed model in terms of the *median* absolute percentage error; however, as 15% of the forecasts have very high (>100%) errors, the MAPE values are very high 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 improves the predictive accuracy significantly relative to Stage 1. Similarly, information sharing – both across stores within a category and across categories within a store – improves the performance significantly relative to the use of the focal residual series only. Finally, both including the marketing variables and using store-specific seasonality terms improve the accuracy of the forecasts significantly.

The rest of the paper is organized as follows. In the next section, we review the relevant strands of the