Lancaster University
Management School

Management Science

Working Paper 2018:04

Retail forecasting: research and practice

Robert Fildes, Lancaster Centre for Marketing Analytics and Forecasting, Lancaster University Management School, UK
Shaohui Ma, School of Business, Nanjing Audit University, China
Stephan Kolassa, SAP Switzerland

The Department of Management Science
Lancaster University Management School
Lancaster LA1 4YX
UK

LUMS home page: http://www.lums.lancs.ac.uk.
Centre home page: http://www.lancaster.ac.uk/lums/research/research-centres--areas/centre-for-marketing-analytics-and-forecasting/

1. R.Fildes@lancaster.ac.uk 2. shaohui.ma@hotmail.com 3. stephan.kolassa@sap.com

Retail forecasting: research and practice

Robert Fildes[1], Lancaster Centre for Marketing Analytics and Forecasting
Department of Management Science, Lancaster University, LA1 1
Shaohui Ma[2], School of Business, Nanjing Audit University, Nanjing, 211815, China
Stephan Kolassa[3], SAP Switzerland, SAP Switzerland
8274 Tägerwilen, Switzerland

Abstract

This paper first introduces the forecasting problems faced by large retailers, from the strategic to the operational, from the store to the competing channels of distribution as sales are aggregated over products to brands to categories and to the company overall. Aggregated forecasting that supports strategic decisions is discussed on three levels: the aggregate retail sales in a market, in a chain, and in a store. Product level forecasts usually relate to operational decisions where the hierarchy of sales data across time, product and the supply chain is examined. Various characteristics and the influential factors which affect product level retail sales are discussed. The data rich environment at lower product hierarchies makes data pooling an often appropriate strategy to improve forecasts, but success depends on the data characteristics and common factors influencing sales and potential demand. Marketing mix and promotions pose an important challenge, both to the researcher and the practicing forecaster. Online review information too adds further complexity so that forecasters potentially face a dimensionality problem of too many variables and too little data. The paper goes on to examine evidence on the alternative methods used to forecast product sales and their comparative forecasting accuracy. Many of the complex methods proposed have provided very little evidence to convince as to their value, which poses further research questions. In contrast, some ambitious econometric methods have been shown to outperform all the simpler alternatives including those used in practice. New product forecasting methods are examined separately where limited evidence is available as to how effective the various approaches are. The paper concludes with some evidence describing company forecasting practice, offering conclusions as to the research gaps but also the barriers to improved practice.

Keywords; retail forecasting; product hierarchies; big data; marketing analytics; user-generated web content; new products; comparative accuracy; forecasting practice.

1. R.Fildes@lancaster.ac.uk 2. shaohui.ma@hotmail.com 3. stephan.kolassa@sap.com

1. Introduction

The retail industry is experiencing rapid developments both in structure, with the growth in on-line business, and in the competitive environment which companies are facing. There is no simple story that transcends national boundaries, with different national consumers behaving in very different ways. For example, in 2017 on-line retailing accounted for 14.8 % of retail sales in the US, 17.6% in the UK but only 3.4% in Italy contrasting with Germany showing a 3.5% increase to 15.1% since 2015 (www.retailresearch.org/onlineretailing.php). But whatever the retailer's problem, its solution will depend in part on demand forecasts, delivered through methods and processes embedded in a forecasting support system (FSS). High accuracy demand forecasting has an impact on organizational performance because it improves many features of the retail supply chain. At the organizational level, sales forecasts are essential inputs to many decision activities in functional areas such as marketing, sales, and production/purchasing, as well as finance and accounting. Sales forecasts also provide the basis for national, regional and local distribution and replenishment plans.

Much effort has been devoted over the past several decades to the development and improvement of forecasting models. In this paper we review the research as it applies to retail forecasting, drawing boundaries around the field to focus on food, non-food including electrical goods (but excluding for example, cars, petrol or telephony), and non-store sales (catalog and now internet). This broadly matches the definitions and categories adopted, for example, in the UK and US government retail statistics. Our objective is to draw together and critically evaluate a diverse research literature in the context of the practical decisions that retailers must make that depend on quantitative forecasts. In this examination we look at the variety of demand patterns in the different marketing contexts and levels of aggregation where forecasts must be made to support decisions, from the strategic to the operational. Perhaps surprisingly, given the importance of retail forecasting, we find the research literature is both limited and often fails to address the retailer's decision context.

In the next section we consider the decisions retailers make, from the strategic to the operational, and the different levels of aggregation from the store up to the retail chain. Section

three considers aggregate forecasting from the market as a whole where, as we have noted, rapid changes are taking place, down to the individual store where again the question of where stores should be located has risen to prominence with the changes seen in shopping behavior. We next turn to more detailed Stock Keeping Unit (SKU) forecasting, and the hierarchies these SKUs naturally fall into. The data issues faced when forecasting include stock-outs, seasonality and calendar events while key demand drivers are the marketing mix and promotions. On-line product reviews and social media are new information sources that requires considerable care if they are to prove valuable in forecasting. Section 5 provides an evaluation of the different models used in product level demand forecasting in an attempt to provide definitive evidence as to the circumstances where more complex methods add value. New product forecasting requires different approaches and these are considered in Section 6. Practice varies dramatically across the retail sector, in part because of its diversity, and in Section 7 we provide various vignettes based on case observation which capture some of the issues retailers face and how they provide operational solutions. Finally, Section 8 contains our conclusions as to those areas where evidence is strong as to best practice and where research is most needed.

2. Retailers' forecasting needs

Strategic level

Retailers like all commercial organizations must make decisions as to their strategic development within a changing competitive and technological environment. The standard elements defining a retail strategy embracing market and competitive factors within the developing technological and regulatory environment (see, for example, Levy, Weitz, and Grewal, 2012) are typically dependent on forecasts. Fig.1 illustrates these issues showing the recent growth of on-line purchases in the US, UK and Europe, with some suggestion that those countries with the highest penetration levels are seeing a slowing of growth (but with clear differences between countries and cultures). Also shown is a naïve extrapolation for 2020 using the average growth rate from 2014 to 2016. Fig.2 shows the changing share of low-price retailers in the UK and the US from 1994 to 2017 with forecasts to 2020 compared to the established leaders (produced via ETS). These simple extrapolative forecasts highlight the

strategic threat on-line and low price retailers pose, exacerbated by a dominant player in Amazon.

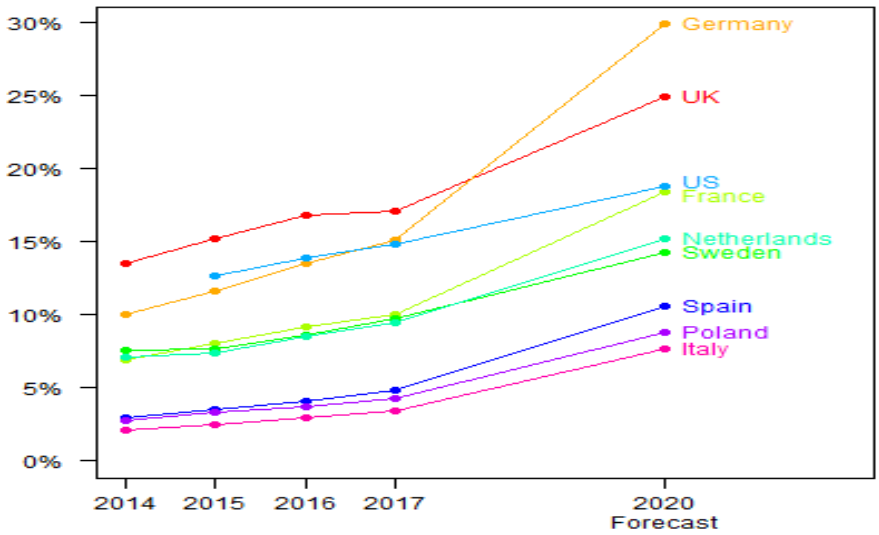

Fig. 1 Online shares of Retail Trade

(Source: Center for Retail Research: www.retailresearch.org/onlineretailing.php)

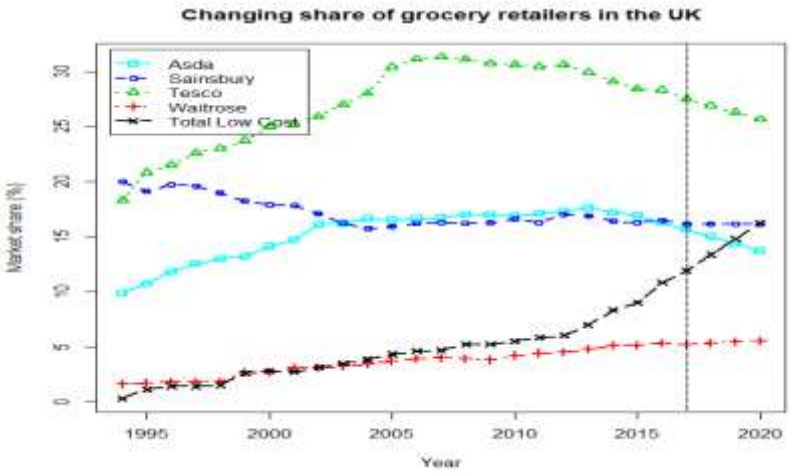

Fig. 2 Share of grocery retailers compared to the low price retailers (Aldi and Lidl) in the UK, 1994 to 2017 with ETS forecasts to 2020. Source: http://www.fooddeserts.org/images/supshare.htm.

These figures and the extrapolative forecasts show the rapid changes in the retail environment which require companies to respond. For example, a channel decision to develop an on-line presence will depend on a forecast time horizon looking decades ahead but with some quantitative precision required over shorter horizons, perhaps as soon as its possible implementation a year or more ahead. The retailer chain's chosen strategy will require decisions that respond to the above changes: on location including channels, price/quality position and target market segment(s), store type (in town vs megastores) and distribution network. A key point is that such decisions will all typically have long-term consequences with high costs incurred if subsequent changes are needed, flexibility being low (e.g. site location and the move to more frequent local shopping in the UK, away from the large out-of-town stores, leading Tesco in 2015 to sell 14 of its earmarked sites in the UK and close down others and, in 2018, M&S proposing to close down more than 10% of its stores). Strategic forecasts are therefore required at both at a highly aggregate level and also a geographic specific level over a long forecast horizon.

The small local retailer faces just as volatile an environment, with uncertainty as to the location and target market (and product mix). Some compete directly with national chains where the issue is what market share can be captured and sustained. But while many of the questions faced by the national retailers remain relevant (e.g. on-line offering) there is little in the research literature that is even descriptive of the results of the many small shop location decisions. Exceptions include charity shops (Alexander, Cryer, and Wood, 2008) and convenience stores (Wood and Browne, 2007) while a number of studies examine restaurants which are outside our scope. But in this article, we focus on larger retailers carrying a wide range of products.

Tactical level

Tactical decisions necessarily fit within the strategic framework developed above. But these strategic decisions do not determine the communications and advertising plan for the chain, the categories of products to be offered, nor the variety (range) of products within each category. At the chain level, the aim is to maximize overall profitability using both advertising (at chain and store level) and promotional tools to achieve success.

At the category level the objective again is to maximize category (rather than brand) profits which will require a pricing/ promotional plan that determines such aspects as the number and depth of promotions over the planning horizon (of perhaps a year), their frequency, and whether there are associated display and feature advertising campaigns. These plans are in principle linked to operational promotional pricing decisions discussed below. The on-shelf availability of products is also a key metric of retail service, and this depends crucially on establishing a relationship between the product demand forecasts, inventory investment and the distribution system. The range of products listed raises the question of new product introduction into a category, the expected sales and its effect on sales overall (particularly within category).

Demands placed on the warehouse and distribution system by store × product demand also need forecasting. This is needed to plan the workforce where the number and 'size' of products determines the pick rate which in turn determines the workforce and its schedule. The constitution of the delivery fleet and planned routes similarly depend on store demand forecasts (somewhat disaggregated) since seasonal patterns of purchasing vary by region. This is true whether the retailer runs its own distribution network or has it outsourced to a service provider – or, what is most common, uses a mixture, with many products supplied from the retailer's own distribution centers, but others supplied directly by manufacturers to stores (Direct Store Delivery).

Operational level

To be successful in strategic and tactical decisions, the retail company needs to design its demand and supply planning processes to avoid customer service issues along with unnecessarily high inventory and substantial write off costs due to obsolete products. These are sensitive issues in retail companies because of the complexity in the demand data with considerable fluctuations, the presence of many intermediaries in the process, diversity of products and the service quality required by the consumer. In a general way, accurate demand forecasting is crucial in organizing and planning purchasing, distribution, and the labor force, as well as after-sales services. Therefore, the ability of retail managers to estimate the probable sales quantity at the SKU × store level over the short-term leads to improved customer satisfaction, reduced waste, increased sales revenue and more effective and efficient

distribution.

As a result of these various operational decisions with their financial consequences, the cash retailers generate (since suppliers are usually paid in arrears) leads to a cash management investment problem. Thus the cash available for investment, itself dependent on the customer payment arrangements, needs to be forecast.

Day-to-day store operations are also forecast dependent. In particular, staffing schedules depend on anticipated customer activity and product intake.

3. Aggregate retail sales forecasting

All forecasting in retail depends on a degree of aggregation. The aggregations could be on product units, location or time buckets or promotion according to the objective of the forecasting activity.

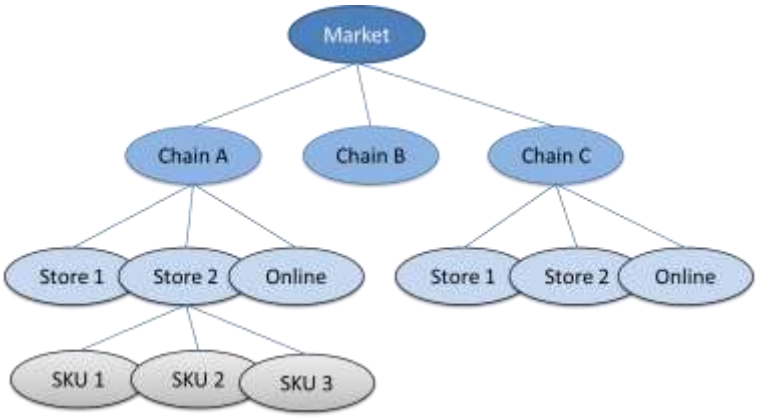

Fig. 3 Hierarchy of aggregate retail sales forecasting

In this section, the aggregate retail sales forecasting refers to the total retail sales in a market, a chain, or a store, as opposed to product (SKU/brand/category) specific forecasts, i.e., we implicitly aggregate across products and promotions and up to a specific granularity (e.g., weekly or monthly) in the time dimension, see Fig.3. Aggregate retail sales are usually measured as a dollar amount instead of units of the products. We below review the existing researches on three levels separately: the aggregate retail sales in a market, in a chain, and in a

store. Though forecasting aggregate sales at these three levels share many common issues, e.g., seasonality and trend, they raise different forecasting questions; have different objectives, data characteristics, and solutions.

3.1 Market level aggregate sales forecasting

Market level aggregate sales forecasting concerns the forecasts of total sales of a retail format, section, or the whole industry in a country or region. The time bucket for the market level forecasts may be monthly, quarterly or yearly. The forecasts of market level retail sales are necessary for (large) retailers both to understand changing market conditions and how these affect their own total sales (Alon, Qi, and Sadowski, 2001). They are also central to the planning and operation of a retail business at the strategic chain level in that they help identify the growth potential of different business modes and stimulate the development of new strategies to maintain market position.

Market level aggregate retail sales data often exhibit strong trend, seasonal variations, serial correlation and regime shifts because any long span in the data may include both economic growth, inflation and unexpected events (**Fig. 4**). Time series models have provided a solution to capturing these stylized characteristics. Thus, time series models have long been applied for market level aggregate retail sales forecasting (e.g., Alon et al., 2001; Bechter and Rutner, 1978; Schmidt, 1979; Zhang and Qi, 2005). Simple exponential smoothing and its extensions to include trend and seasonal (Holt-Winters), and ARIMA models have been the most frequent time series models employed for market level sales forecasting. Even in the earliest references, reflecting controversies in the macroeconomic literature, the researchers raised the question of which of various time series models performed best and how they compared with simple econometric models[1]. The early studies suffered from a common weakness – a failure to compare models convincingly.

[1] Typically, macro econometric models do not include retail sales as an endogenous variable but rather use a variable such as consumption.

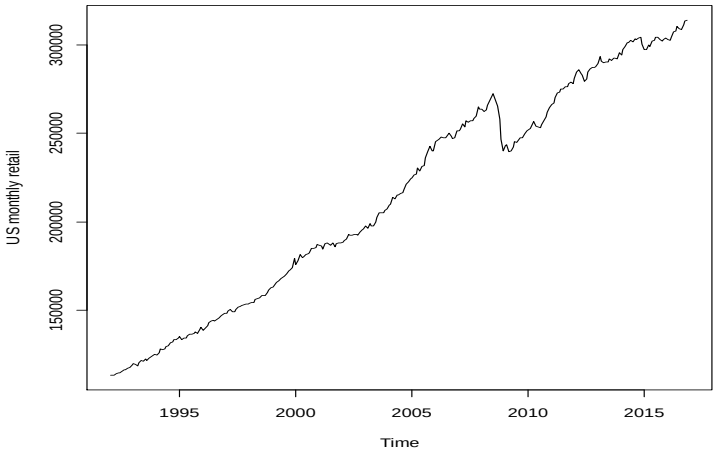

Fig. 4 US retail sales monthly series in million dollars.

(Source: U.S. Census Bureau)

Some researchers found that standard time series models were sometimes inadequate to approximate aggregate retail sales, identifying evidence of nonlinearity and volatility in the market level retail sales time series. Thus, researchers have resorted to nonlinear models, especially artificial neural networks (Alon, et al., 2001; Chu and Zhang, 2003; Zhang and Qi, 2005). Results have indicated that traditional time series models with stochastic trend, such as Winters exponential smoothing and ARIMA, performed well when macroeconomic conditions were relatively stable. When economic conditions were volatile (with rapid changes in economic conditions) ANNs was claimed to outperform the linear methods (Alon et al., 2001) though there must be a suspicion of overfitting. One study also found that prior seasonal adjustment of the data can significantly improve forecasting performance of the neural network model in forecasting market level aggregate retail sales (Kuvulmaz, Usanmaz, and Engin, 2005) although in wider NN research this conclusion is moot. Despite these claims this evidence of the forecasting benefits of non-linear models seems weak as we see below.

Econometric models depend on the successful identification of predictable explanatory variables compared to the time series model. Bechter and Rutner (1978) compared the forecasting performance of ARIMA and econometric models designed for US retail sales. They used two explanatory variables in the economic model: personal income and nonfinancial

personal wealth as measured by an index of the price of common stocks; past values of retail sales were also included in alternative models that mixed autoregressive and economic components. They found that ARIMA forecasts were usually no better and often worse than forecasts generated by a simple single-equation economic model, and the mixed model had a better record over the entire 30-month forecast period than any of the other three models. No ex ante unconditional forecast comparisons have been found. Recently, Aye, Balcilar, Gupta, and Majumdar (2015) conducted a comprehensive comparative study over 26 (23 single and 3 combination) time series models to forecast South Africa's aggregate retail sales. Unlike the previous literature on retail sales forecasting, they not only looked at a wide array of linear and nonlinear models, but also generated multi-step-ahead forecasts using a real-time recursive estimation scheme over the out-of-sample period. In addition, they considered loss functions that overweight the forecast error in booms and recessions. They found that no unique model performed the best across all scenarios. However, combination forecast models, especially the discounted mean-square forecast error method (Stock and Watson, 2010) which weights current information more than past, not only produced better forecasts, but were also largely unaffected by business cycles and time horizons.

In summary, no research has been found that uses current econometric methods to link retail sales to macroeconomic variables such as GDP and evaluate their conditional and unconditional performance compared to time series approaches. The evidence on the performance of non-linear models is limited with too few series from too few countries and the comparison with econometric models has not been made.

3.2 Chain level aggregate sales forecasting

Research at the retail chain level has mainly focused on sales forecasting one year-ahead (Curtis, Lundholm, and McVay, 2014; Kesavan, Gaur, and Raman, 2010; Osadchiy, Gaur, and Seshadri, 2013). Accurate forecasts of chain level retail sales (in money terms) are needed for company financial management and also to aid financial investment decisions in the stocks of retail chains.

In general, most of the models used for chain level are similar to those used for market

level forecasting (i.e. univariate extrapolation models). However, there are some specially designed models which have been found to have better performance. Kesavan et al. (2010) found that inventory and gross margin data can improve forecasting of annual sales at the chain level in the context of U.S. publicly quoted retailers. They incorporated cost of goods sold, inventory, and gross margin (the ratio of sales to cost of goods sold) as endogenous variables in a simultaneous equations model, and showed sales forecasts from this model to be more accurate than consensus forecasts from equity analysts. Osadchiy et al. (2013) presented a (highly structured) model to incorporate lagged financial market returns as well as financial analysts' forecasts in forecasting firm-level sales for retailers. Their testing indicated that their method improved upon the accuracy of forecasts generated by equity analysts or time-series methods. Their use of benchmark methods (in particular a more standard econometric formulation) was limited. Building on earlier research Curtis et al. (2014) forecast retail chain sales using publicly available data on the age mix of stores in a retail chain. By distinguishing between growth in sales-generating units (i.e., new stores) and growth in sales per unit (i.e., comparable store growth rates), their forecasts proved significantly more accurate than the forecasts from models based on estimated rates of mean reversion in total sales as well as analysts' forecasts. Internal models of chain sales forecasts should benefit from including additional confidential variables but no evidence has been found.

3.3 Store level aggregate sales forecasting

Retailers typically have multiple stores of different formats, serving different customer segments in different locations. Store sales are dramatically impacted by location, the local economy and competitive retailers, consumer demographics, own or competitor promotions, weather, seasons and local events including for example, festivals. Forecasting store sales can be classified into two categories: (1) forecasting existing store sales for distribution, target setting and viability, and financial control, and (2) forecasting new store potential sales for site selection analysis.

Both univariate time series and regression models are used for forecasting existing store sales. Steele (1951) reported on the effect of weather on the daily sales of department stores.

Davies (1973) used principal components and factor analysis in a clothing-chain study and demonstrated how the scores of individual stores on a set of factors may be interpreted to explain their sales performance levels. Geurts and Kelly (1986) presented a case study of forecasting department store monthly sales. They considered various factors in their test models including seasonality, holiday, number of weekend days, local consumer price index, average weekly earnings, and unemployment rate, etc. They concluded that univariate time series methods were better than judgment or econometric models at forecasting store sales. At a more operational level of managing staffing levels, Lam, Vandenbosch, and Pearce (1998) built a regression model based on daily data which set store sales potential as a function of store traffic volume, customer type, and customer response to sale force availability: the errors are modelled as ARIMA processes. However, no convincing evidence was presented on comparative accuracy. With the rapid changes on the high-street in many countries showing increasing vacancy rates, these forecasting models will increasingly have a new use: to identify shops to be closed. We speculate that multivariate time series models including indicator variables (for the store type), supplemented by local knowledge, should prove useful. But this is research still to be done.

Forecasting new store sales potential has been a difficult task, but crucial for the success of every retailing company. Traditionally, new store sales forecasting approaches could be classified into three categories: judgmental, analogue regression and space interaction models (also called gravitational models). Note that any evaluation of new store forecasts needs to take a potential selection bias into account: candidate new stores with higher forecasts are more likely to be developed and may see systematically lower sales than forecasted because of regression to the mean. (The analogue is also true for forecasting new *product* sales or promotional sales, see below.)

The success of the judgmental approach depends on the experience of the location analyst (Reynolds and Wood, 2010). Retailers often use the so-called “checklist” to systematically assess the relative value of a site compared to other potential sites in the area. It can deal with issues that cannot be expressed quantitatively (e.g. access; visibility) and is where intuition and experience become important. In its simplest form the checklist can act as a good screening tool

but is unable to predict turnover. The basic checklist approach can be further developed to emphasize “some variable points rating” to factors specific to success in particular sectors, for example, convenience store retailing (Hernandez and Bennison, 2000).

The analogue regression generates turnover forecasts for a new store by comparing the proposed site with existing analogous sites, measuring features such as competition (number of competitors, distance to key competitor, etc.), trading area composition (population size, average income, the number of households, commute patterns, car ownership, etc.), store accessibility (cost of parking, distance to parking, distance to bus station, etc.) and store characteristics (size, format, brand image, product range, opening hours, etc.). Compared with the judgmental approach, analogue regression models provide a more objective basis for the manager's decision-making, highlighting the most likely options for new locations. Simkin (1989) reported the success application of a regression based Store Location Assessment Model (SLAM) in several of the UK's major retailers. The model was able to account for approximately 80% of the store turnover, but prediction accuracy for the sales of new stores is not reported in the paper. Morphet (1991) applied regression to an analysis of the trading performance of a chain of grocery stores in the England incorporating five competitive and demographic factors (including population, share of floor space, distance higher order centre, pull, percentage of married women, etc.). Though the models achieved a high degree of 'explanation' of the variation in store performance, the results on predicted turnover suggested that the use of regression equations was insufficient to predict the potential performance of stores in new locations. The pitfalls of regressions may come from statistical overfitting due to limited data, neglecting consumer perceptions, and inadequate coverage of competition. While the method can include various demographic variables and is therefore appropriate for retail operations aiming for a segmented market it is heavily data dependent and therefore of limited value for a rapidly changing retail environment (as in the UK).

The spatial interaction model (SIM) (or gravity model) is a widely used sophisticated retail location analysis tool, which has a long and distinguished history in the fields of geography and regional science. Based on Reynold and Wood’s (2010) survey of corporate location planning departments, around two thirds of retail location planning teams (across all sectors) make use

of SIM for location planning. Different from analogous regressions which mainly rely on the data from existing stores in the same chain, SIM uses data from various sources to improve prediction accuracy: analogous stores, household surveys, geographical information systems, competition and census data. A spatial interaction model is based on the theory that expenditure flows and subsequent store revenue are driven by the store's potential attractiveness and constrained by distance, with consumers exhibiting a greater likelihood to shop at stores that are geographically proximate (Newing, Clarke, and Clarke, 2014). The basic example of this type of model is the Huff trade area model (Huff, 1963). Its popularity and longevity can be attributed to its conceptual appeal, relative ease of use, and applicability to a wide range of problems, of which predicting consumer spatial behavior is the most commonly known (Li and Liu, 2012). The original Huff model has been extended by adding additional components to make the model more realistic; these include models that can take into account retail chain image (Stanley and Sewall, 1976), asymmetric competition in retail store formats (Benito, Gallego, and Kopalle, 2004), store agglomeration effects (Li and Liu, 2012; Picone, Ridley, and Zandbergen, 2009; Teller and Reutterer, 2008), retail chain internal cannibalization (Beule, Poel, and Weghe, 2014), and consumer heterogeneity (Newing, et al., 2014). Furthermore, spatial data mining techniques and GIS simulation have been applied in retail location planning. These new techniques have proved to outperform the traditional modeling approach with regard to predictive accuracy (Lv, Bai, Yin, and Dong, 2008; Merino and Ramirez-Nafarrate, 2016).

Following Newing et al. (2014), let S_{ij} represent the expenditure flowing between zone i and store j then

$$S_{ij} = O_i \frac{W_j \exp(-\beta C_{ij})}{\sum_j W_j \exp(-\beta C_{ij})}$$

O_i is a measure of the demand (or expenditure available in zone i); C_{ij} represents the travel time between zone i and store j; and W_j accounts for the attractiveness of store j. The attractiveness term, W_j will itself depend on factors such as accessibility, parking, other store features etc. Such models are usually validated on in-sample data. But Birkin, Clarke, and Clarke (2010) criticize this limited approach emphasizing the importance of a hold-out sample (an unacknowledged reference to the forecasting literature) and show, using DIY chain store data,

that the model can be operationalized with a forecasting accuracy of around 10% (which proved better than the company's performance). An important omission is the time horizon over which the model is assumed to apply, presumably the time horizon of the investment. Birkin et al. (2010) comment the models are regularly updated at least annually which suggests an implicit view as to lack of longer-term stability in the models arising from a changing retail environment. Extensions to the model suffer from problems of data inadequacies but Newing et al. (2014) argue these can be overcome to include more sophisticated demand terms such as seasonal fluctuation，and different types of retail consumer with different shopping behaviors.

Predictive models of store performance are only one element in supporting the location decision. Wood and Reynolds (2013) discuss how the models are combined with context specific knowledge and the judgments of location analysts and analogous information to produce final recommendations. There is no evidence available on the relative importance of judgmental inputs and model based information. Nor is there much evidence on the accuracy of the models beyond untested claims as to the model based forecasts being highly accurate (Wood and Reynolds, 2013) apart from Birkin et al.'s (2010) analysis of a DIY chain. In the rapidly changing retail environment, we speculate that judgment will again become the dominant approach to evaluating store potential and store closures. The research question now becomes what role if any models can usefully play.

Short-term forecasting of store activity can utilize recently available 'big' data in the form of customer credit (or mobile) transactions to produce shop sales forecasts. The use of the forecasts a week or so ahead is in staff scheduling. Ma and Fildes (2018) used mobile sales transactions, aggregated to daily store level for 2000 shops registered on a leading third-party mobile payment platform in China to show that the forecasts which took into account the overall activity on the platform (i.e. a multivariate approach) produced using a machine learning algorithm, outperformed univariate methods including standard benchmarks.

4. Product level demand forecasting in retail

Product level demand forecasting in retail usually aims to generate forecasts for a large number of time series over a short forecasting horizon, in contrast to long term forecasting for

only one or a few of time series at a more aggregate level. The ability to accurately forecast the demand for each item sold in each retail store is critical to the survival and growth of a retail chain because many operational decisions such as pricing, space allocation, availability, ordering and inventory management for an item are directly related to its demand forecast. Order decisions need to ensure that the inventory level is not too high, to avoid high inventory costs, and not too low to avoid stock out and lost sales.

4.1 The hierarchical structure of product level demand forecasting

In general, given a decision-making question, we then need to characterize the product demand forecasting question on three dimensions: the level in the product hierarchy, the position in the retail supply chain, and the time granularity (**Fig. 5**): these are sometimes labelled 'data cubes'..

Fig. 5 Multidimensional hierarchies in retail forecasting

Time granularity

For different managerial decisions, demand forecasts are needed at different time granularities. In general, the higher the level of the decision from the operational to the strategic, the lengthier the forecasting time granularity. For example, we may need forecasts on daily granularity for store replenishment, on a weekly level for DC replenishment,

promotion planning, and (initial) allocation planning, while on-line fashion sales may rely on an initial estimate of total seasonal sales, updated just once mid-season.

Product aggregation level

Three levels of the product hierarchy are often used for planning by retailers: SKU level, brand level, and category level.

SKU is the smallest unit for forecasting in retail, which is the basic operational unit for planning daily stock replenishment, distribution and, promotion. SKU level forecasts are usually conducted across stores up to the chain as a whole and in daily/weekly time steps. The number of SKUs in a retail chain may well be huge. E.g., in a supermarket, drugstore or home improvement/do it yourself (DIY) retailer today, tens thousands of items need weekly or even daily forecasts. Walmart faces the problem of over one billion SKU × Store combinations (Seaman, 2018). In a fashion chain such as Zara the number of in-store items by design, colour and size can also be of the order of tens of thousands, although forecasting may be conducted at the "style" or design level, aggregating historical data across sizes and colours and disaggregating using size curves and proportions to arrive at the final SKU forecasts. Online assortments are typically far larger, especially in the fashion, DIY or media (books, music, movies) business.

A brand in a product category often includes many variant SKUs with different package types, sizes, colors, or flavors. In addition to SKU level promotional planning, brand level forecasts are also important where there are cross-brand effects and promotions and ordering may be organized by brand.

However, for many retail decisions, the initial forecasts that are required are more aggregate, with a tactical promotional plan being developed across the chain that may well take inter-category constraints into account (although whether in practice forecasts have an active role in such a plan is an open question). A product category usually contains tens of brands or hundreds of SKUs with certain attributes in common, e.g., canned soup, shampoo or nails. Categories may be segmented into subcategories, which may be nested in or cut across brands. Category level sales forecasting mainly focuses on weekly or monthly forecasts in a store, over a chain or over a market, and such forecasts are mainly used for budget planning

by so-called category managers, who make large scale budgeting, planning and purchasing decisions, which again need to harmonize with the resources needed to actually execute these decisions, e.g., shelf space, planograms or specialized infrastructure like available freezer space.

Category management and the assortment decision starts with a category forecast which Kök, Fisher, and Vaidyanathan (2015) suggest is based on trend analysis supplemented by judgment. The assortment decision on which brands (or SKUs) to exclude as well as which new products to add is dependent on the SKU level demand forecasts: the effects on aggregate category sales of the product mix depend on the cross-elasticities of the within category SKU level demand forecasts, with a long (12 month) time horizon. The associated shelf-allocation is, Borin and Farris (1995) claim, insensitive to SKU demand forecast errors.

In short, whatever the focus, SKU level forecasts as well as their associated own and cross-price elasticities are needed to support both operational and tactical decisions.

Supply Chain

A typical retail supply chain consists of manufacturers, possibly wholesalers or other intermediaries, retailers' distribution centers (DCs), and stores in different formats. Retailers need forecasts for the demands faced by each level in the supply chain. Product-store level forecasting is often for replenishment, product-DC level forecasting for distribution, product-chain level forecasting for preordering, brand-chain level for supplier negotiations and potentially for manufacturing decisions in vertically integrated retailers, such as increasingly many fashion chains. A key question in retail supply chain forecasting is how to collaborate and integrate the data from different supply chain levels so that forecasts at different levels of the supply chain are consistent and provide the required information to each single decision-making process. From the retailer's perspective the coordination whilst costly has the potential to improve availability and lower inventory. It may improve retail forecasting accuracy or service levels (Wang and Xu, 2014) though some retailers doubt this, apparently only selling rather than sharing their data. Empirical models analyzing the relationship between POS data and manufacturing forecast accuracy show improvements are possible though not inevitable (Hartzel and Wood, 2017;Trapero, Kourentzes, and Fildes, 2012; Williams, Waller, Ahire, and

Ferrier, 2014). Empirical evidence on successful retail implementation is limited though Smaros (2007) using case studies identified some of the barriers and how they might be overcome (Kaipia, Holmström, Småros, and Rajala, 2017).

4.2 Forecasting within a product hierarchy

Given a specific retail decision-making question, we first need to determine the aggregation level for the output of the sales forecasting process. A common option is to choose a consistent level of aggregation of data and analysis. For example, if one needs to produce demand forecasts at the SKU-weekly-DC level it might seem ‘‘natural’’ to aggregate sales data to the SKU-weekly-DC level and analyze them at the same level as well. However, the forecasts can also be made by two additional forecasting processes within the data hierarchy: (1) the bottom-up forecasting process and (2) the top-down forecasting process.

The choice of the appropriate level of aggregation depends on the underlying demand generation process. Existing researches have shown that the bottom-up approach is needed when there are large differences in structure across demand time series and underlying drivers (Orcutt and Edwards, 2010; Zellner and Tobias, 2000; Zotteri and Kalchschmidt, 2007; Zotteri, Kalchschmidt, and Caniato, 2005). This is particularly true when the demand time series are driven by item specific time-varied promotions. Foekens, Leeflang, and Wittink (1994) found that disaggregate models produce higher relative frequencies of statistically significant promotion effects with magnitudes in the expected ranges. However, in the case of many homogeneous demand series and small samples, the top-down approach can generate more accurate forecasts (Jin, Williams, Tokar, and Waller, 2015; Zotteri and Kalchschmidt, 2007; Zotteri et al., 2005). For instance, different brands of ice cream will have a similar seasonality with a summer peak, which may not be easily detected for low-volume flavors but can be estimated at a group level and applied on the product level (Syntetos, Babai, Boylan, Kolassa, and Nikolopoulos, 2016). Song (2015) suggested that it is beneficial to model and forecast at the level of data where stronger and more seasonal information can be collected.

In order to solve the trade-off, cluster analysis has been found useful in improving the forecast performance (Boylan, Chen, Mohammadipour, and Syntetos, 2014; Chen and Boylan,

2007). For example, when aggregating product category level demand over stores, one can cluster stores according to whether they have similar demand patterns rather than according to their geographical proximity. A priori clustering based on store characteristics such as size, range and location is common. Appropriately implemented clustering can enable the capture of differences among stores (e.g., in terms of price sensitivity) as the clustering procedure groups stores with similar demand patterns (e.g., with similar reaction to price changes). In these terms, clustering is capable of resolving the trade-off between aggregate parameterization and heterogeneity, leading towards more efficient solutions. But so far, the weight of contributions on this issue focused only on the use of aggregation to estimate seasonality factors (Chen and Boylan, 2007). These works provided evidence that aggregating correlated time series can be helpful to better estimate seasonality since it can reduce variability.

Hyndman, Ahmed, Athanasopoulos, and Shang (2011) proposed a method for optimally reconciling forecasts of all series in a hierarchy to ensure they add up consistently over the hierarchy levels. Forecasts on all-time series in the hierarchy are generated separately first and these separate forecasts are then combined using a linear transformation. So far the approach has not been examined for retail demand forecasting applications.

In general, hierarchical forecasting has received significant attention, but most researchers consider only the aggregation problem for general time series, and have not considered the characteristics of retail sales data which are affected dramatically by many common factors, such as events, promotions and weather conditions. Research by Jin et al. (2015) suggests that for store×SKU demand, in promotional intensive categories, regression based methods including many of the factors discussed above produce substantially more accurate forecasts. At higher levels of aggregation, in time and space, time series methods may well be adequate (Weller, Crone, and Fildes, 2016) though research for retail data remains to be done. But there is as yet no straightforward answer as to how to generate consistent demand forecasts on multiple hierarchies over different dimensions.

4.3 Product level retail sales data characteristics and the influential drivers of demand

At the product level, many factors may affect the characteristics of the observed sales data and underlying demand. Some of the factors are within the control of retailers (such as pricing and promotions, and "secondary" effects like interaction or cannibalization effects from listed, delisted or promoted substitute or complementary products), other factors are not controllable, but their timing is known (such as sporting events, seasons and holidays), and some factors are themselves based on forecasts (such as the competition, local and national economy and weather). There are also many other unexpected drivers of retail sales, such as abnormal events (like terror attacks or health scares), which manifest themselves as random disturbances to sales time series which are correlated across category and stores that share common sensitive characteristics.

As the result of these diverse effects, product level sales data are characterized by high volatility and skewness, multiple seasonal cycles, their often large volume, intermittence with zero sales frequently observed at store level, together with high dimensionality in any explanatory variable space. In addition, the data are also contaminated by stock-outs where the consumer is unable to purchase the product desired and instead may shift to another brand or size or, in the extreme, leave to seek out a related competitor.

Stock-outs: demand vs. sales

Retail product level demand forecasting usually depends on the SKU sales data typically captured by POS transactions. However, POS sales data presents an imperfect observation of true demand due to the demand censoring effect, when the actual demand exceeds the available inventory. Demand estimates using only sales data would result in a negative bias in demand estimates of the focal product. At the same time, customers may turn to purchase substitutes when facing a stock-out in the primary target product: this may increase the sales of substitute products and result in an overestimate of the substitutes. Academic researchers have long recognized the need to account for this censoring effect in inventory management. This literature has been primarily centered on methodologies for dealing with the imperfect demand

observations. The methods can be classified into two categories: nonparametric (e.g., Kaplan and Meier,1958) and parametric models using hazard rate techniques (e.g., Wecker, 1978; Nahmias, 1994; Agrawal and Smith, 1996). For more detail, see Tan and Karabati (2004) who provided a review on the estimation of demand distributions with unobservable lost sales for inventory control. Most of methods are based on stock out events data, while Jain, Rudi, and Wang (2014) found that stock-out timing could further improve the estimation accuracy compared with methods based on stock-out events. In the marketing and assortment management literatures, researchers have focused on the consumers' substitution seeking behavior when their target product is facing stock out, which is another way of viewing the problem of product availability (e.g., Kök and Fisher , 2007; Vulcano, Ryzin, and Ratliff , 2012; Conlon and Mortimer, 2013).

Conversely, there is some evidence that at least for some categories, demand depends on inventory, with higher inventory levels driving higher sales: this has been called a "billboard effect" (Koschat, 2008; Ton and Raman, 2010). Anecdotally, we have encountered retailers who know this putative effect as "product pressure". However, no literature appears to have leveraged inventories as a driver to improve forecasts.

The proposed forecasting models in this area are in general explanatory and often require more information than is readily available, such as periodic stock auditing, customer numbers and assortment information. In addition, any forecasting algorithm that leverages system inventory information needs to deal with the fact that system inventories are notoriously inaccurate (so-called "Inventory Record Inaccuracy" or IRI (Dehoratius and Raman, 2008). As a consequence models published so far are not suited to forecasting applications. The limited research reported in the forecasting literature may in part be due the lack of real demand observations so forecasting accuracy is hard to measure. On the other hand, storing observed changes in the shelf inventory for every product may be very costly to the retailer, and may not be adequate to identify every single stock-out instance. Technological solutions may become more common such as RFID (Bottani, Bertolini, Rizzi, and Romagnoli, 2017). The forecasting issue is whether out-of-stock positions affect overall service and profitability (within category).

Intermittence

Intermittence is another common characteristic in store POS sales data, especially in slow moving items at daily SKU level. **Fig**.6 depicts a SBC (Syntetos, Boylan, and Croston, 2005) categorization (see also Kostenko and Hyndman, 2006) over the daily sales of 1373 household cleaning items from a UK retailer, cross-classified by the coefficient of variation in demand and the mean period between non-zero sales. 861 items exhibit strong intermittent characteristics.

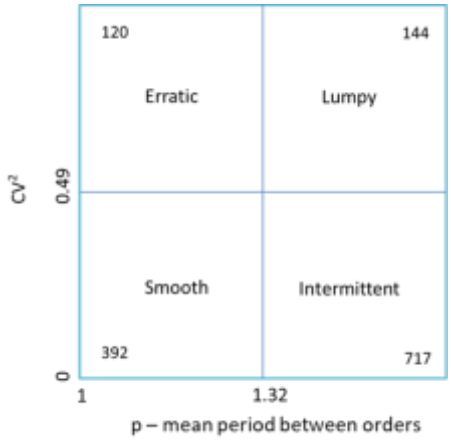

Fig. 6 SBC categorization on 1373 household clean items (Source: UK supermarket data)

Techniques designed specifically for intermittent demand include Croston's method (Croston, 1972), the Syntetos and Boylan method (Syntetos and Boylan, 2001), Levén and Segerstedt method (Levén and Segerstedt, 2004), Syntetos–Boylan approximation (SBA) method (Syntetos and Boylan, 2005), and TSB method (Teunter, Syntetos, and Zied Babai, 2011), etc. However, most of these models are tested on demand/ sales time series data from industries other than retail (e.g., service/spare parts, high-priced capital goods in electronics, automotive, aerospace and high tech), except for Kolassa (2016), who assessed density forecasts based on Croston's method and found them sorely lacking. Also note that while Croston's method is intuitively appealing and commonly used in practice – at least as a

benchmark –, Shenstone & Hyndman (2005) point out that any possible underlying model will be inconsistent with the properties of intermittent demands, exhibiting non-integer and/or negative demands. Nevertheless, Shenstone & Hyndman note that Croston's point forecasts and prediction intervals may still be useful.

As mentioned in the stock-out discussion, POS sales are not the same as the latent demand. The observed zero sales may either be due to the product's temporary unavailability (e.g., stock out or changes in assortment) or intermittent demand. Without product availability information, it is hard to infer the latent demand using only sales data. Much of the retail forecasting literatures when dealing with forecasting of slow moving items has not recognized this problem in their empirical studies (e.g., Cooper, Baron, Levy, Swisher, and Gogos, 1999; Li and Lim , 2018), while the only exception found is Seeger, Salinas, and Flunkert (2016) who treat demand in a stock-out period (assuming stock-out is observable) as latent in their Bayesian latent state model of on-line demand for Amazon products.

Product level demand in retail is also disturbed by a number of exogenous factors, such as promotions, special events, seasonalities and weather, etc. (as will be discussed in what follows): all of these factors make intermittent demand models difficult to be applied to POS sales data. One possibility is to model these influences on intermittent demands via Poisson or Negative Binomial regression. Kolassa (2016) found that the best models included only day of week patterns. One alternative approach, yet to be explored in retail, is the use of time series aggregation through MAPA (Kourentzes, Petropoulos, and Trapero, 2014) to overcome the intermittence, which then could be translated into distribution centre loading.

Seasonality

Retail product sales data have strong seasonality and usually contain multiple seasonal cycles of different lengths. For example, beer daily sales data shown in 7 exhibit both weekly and annual cycles. Sales are high during the weekends and low during the weekdays, high in summer and low in winter, and high around Christmas. Some sales data may also possess biweekly or monthly (paycheck effects) or even quarterly seasonality, depending on the nature of the business and business locations. For this reason, models used in forecasting must be able to handle multiple seasonal patterns. Ramos and Fildes (2018) demonstrate this point, using

models with sufficient flexibility but parsimonious complexity to capture the seasonality of weekly retail data: trigonometric functions prove sufficient.

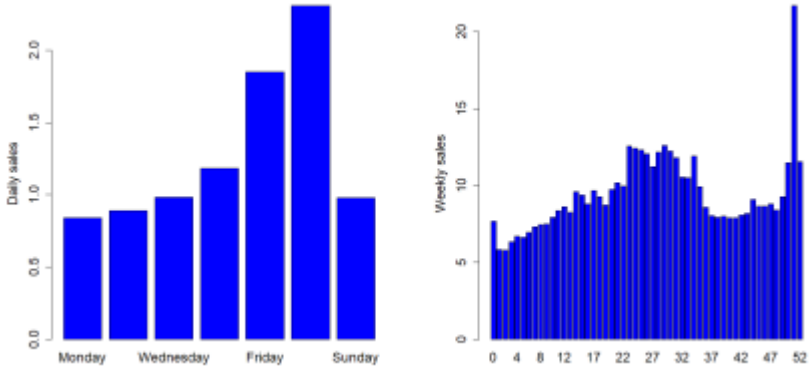

Fig. 7 Beer daily and weekly sales: UK supermarket data

Calendar events

Retail sales data are strongly affected by some calendar events. These events may include holidays (Fig. shows a significant lift in Christmas, i.e., week 51), festivals, and special activities (e.g., important sport matches or local activities). For example, Divakar, Ratchford, and Shankar (2005) found that during holidays the demand for beverages increased substantially, while other product groups were negatively affected. In addition, SKU × Store consumption may change due to changes in the localized temporary demographics. Most research includes dummy variables for the main holidays in their regression models (Cooper et al., 1999). Certain holidays recur at regular intervals and can thus be modeled as seasonality, e.g., Christmas or the Fourth of July in the US. Other holidays move around more or less widely in the (Western style) calendar and are therefore not be captured as seasonality, such as Easter, Labor Day in the US, or various religious holidays whose date is determined based on non-Western calendars, such as the Jewish or the Muslim lunar calendars.

Weather

The demand for some retail products is also strongly affected by temperature and other weather conditions. For example, there is usually strong support that the sales of soft drinks are

higher when the weather is hot (e.g., Cooper et al., 1999; Dubé, 2004). Murray and Muro (2010) found that as exposure to sunlight increases, consumer spending tends to increase. Nikolopoulos and Fildes (2013) showed how a brewing company's simple exponential smoothing method for in-house retail SKU sales could be adjusted (outside the base statistical forecasts) to take into account temperature effects.

Weather effects may well be non-linear. For instance, sales of soft drinks as a function of temperature will usually be flat for low to medium temperatures, then increase with hotter weather, but the increase may taper off with extreme heat, when people switch from sugary soft drinks to straight water. Such effects could in principle be modeled using spline transformations of temperature.

One challenge in using weather data to improve retail sales forecasts is that there is a plethora of weather variables available from weather data providers, from temperatures (mean temperature during a day, or maximum temperature, or measures in between) to the amount, duration and type of precipitation, or the sunshine duration, wind speed or wind chill factors, to even more obscure possibilities. One can either choose some of these variables to include in the model, or transform them in an appropriate way. For instance, one can define a Boolean "barbecue predictor", which is TRUE whenever, say, the temperature exceeds 20 degrees Celsius and there is less than 20% cloud cover. In addition, there are interactions between the weather and other predictors, like promotions or the time of year: sunny weather will have a stronger impact on a promoted ice cream brand than on an unpromoted one, and "barbecue weather" will have a stronger impact on steak sales at the beginning of the summer, when people can observe "the first barbecue of the season", than later in the year after they have been barbecuing for months.

Another hurdle is, of course, that weather variables need to be forecasted themselves, in contrast to intervention variables like prices or promotions that the retailer sets themselves, or calendar events whose date is known with certainty. This means that weather data can only be meaningfully used for short-range sales forecasts, since weather forecasts are better than chance only for a short horizon, or for cleaning *past* data of historical impacts of, say, heat waves. In addition, this aspect implies that forecasting exercises that use the actual weather in *ex post*

forecasting appraisal overstate the forecast's certainty, since they do not include the uncertainty inherent in the weather forecast. This uncertainty can in principle be surmounted in analyzes; however, it has been our experience that historical weather *forecasts* are much more expensive to obtain from data providers than historical *actual weather data*. Plus, one needs to ensure the correct vintage of forecasts: to calculate two-day-ahead sales forecasts, we need two-day-ahead weather forecasts, for three-day-ahead sales forecasts, we need three-day-ahead weather forecasts and so on.

Marketing mix and promotions

The regular price and relative price discount are important variables that should be included in any forecasting model. Where models are focused on short-term forecasts, the promotional price is important. Apart from price effects, the effects of feature advertisement have been studied extensively in the marketing literature. Feature advertisement can be divided into in-store advertisement and other advertisement like using a newspaper or a store flyer to increase store traffic (Gijsbrechts, Campo, and Goossens, 2003). In-store advertisement focuses on attracting customers to the promoted articles in several ways, the most commonly used methods making use of ads and displays (Cooper et al., 1999). The effects of displays have also been intensively investigated in the marketing literature. A general conclusion is that sales can increase several fold in the presence of displays (Ailawadi, Harlam, César, and Trounce, 2006).

Fig.8, which is a result of data exploration of the IRI retail data (Bronnenberg, Kruger, and Mela, 2008), shows that the different marketing instruments have different magnitudes of lift effects on various product categories. Categories with long shelf lives lend themselves to stockpiling or "pantry loading", resulting in stronger promotional uplifts and post-promotional dips than for products with shorter shelf lives, like fresh or ultra-fresh categories. In addition, different promotional types (e.g. Buy-one-get-one-free: BOGOF versus 50% price reduction) have effects not captured by just the unit price, an aspect recognized in some retail software products such as SAP. Finally, marketing instruments may be used in combination: a BOGOF may be announced by shelf tags, on store signage, in the retailer's app, in the newspaper or in some or all of these venues; it may be available to all comers, only to customers using an in-app coupon, or only to loyalty card holders (of a certain tier); it may offer additional loyalty

points or an additional coupon; to get the free unit, the shopper may need to buy one ("vanilla" BOGOF), two or more units; and the promotion may run for one day, one week, or multiple weeks (with typically declining impact); or it may be part of a larger advertising campaign like an "advent calendar". The possible combinations will interact in different ways, and the quest for novelty on the part of retailer marketers ensures that products will regularly need to be forecasted with a marketing mix not previously observed for this particular product – necessitating a kind of "new promotion" forecasting analogous to a degree to new store and new product forecasting.

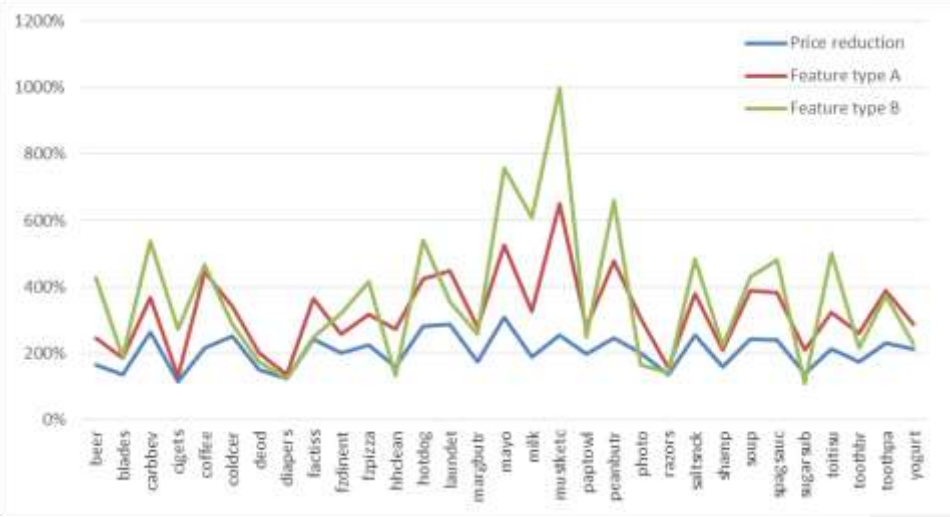

Fig. 8 Promotional lift effects in IRI dataset among various categories. Source: IRI data set.

The majority of the promotional response stems from brand switching or accelerating or "pulling forward" future purchases in the store (Bucklin and Siddarth, 1998; Chiang, 1991; Chintagunta, 1993; Gupta, 1988). That means the promotion on one item may affect the sales of another or its own later sales, so-called "cannibalization". A large body of research supports the view that brands within a product category are substitutes for one another (Kumar and Leone, 1988; Moriarty, 1985; Mulhern and Leone, 1991; Walters, 1988, 1991), and incorporating cannibalization effects can substantially improve forecasts (Srinivasan, Ramakrishnan, and Grasman, 2005), at least at aggregate level. However, it is unclear whether the cannibalization

signal at SKU × store level is strong enough to improve forecasts at this granular level and to actually improve the stock position on the shelf. In addition, modeling cannibalization requires a significant additional effort to identify drivers and victims, although product hierarchies may help here, and a retailer would likely restrict the modeling of cannibalization with its effort to important categories.

Finally, much the same discussion as for cannibalization applies to interaction effects from complementary products. For instance, a promotion on steaks may be hypothesized to increase the sales of steak sauces. The analysis of the two types of interaction – cannibalization and complementarity – differs in two key aspects. First, as noted above, product categories typically group similar products that are likely substitutes, so as noted above, the product category can be used to identify interacting pairs or groups of products. Conversely, the product hierarchy typically does not group complementary articles together, so it is not useful for identifying pairs or groups of products that may exhibit complementarity useful for forecasting. Second, however, basket analysis, i.e., the analysis of transaction log data with a view to which products were bought by the same shopper at the same time, can be useful in detecting complements, using affinity analysis – but basket analysis is harder to leverage to detect substitutes. However, cross-category price elasticities appear to be small, limiting the scope for improving forecasts using complementarity (Russell and Petersen, 2000). Ma, Fildes, and Huang (2016) show that improvements of 12.6% in forecasting accuracy (as measured by Mean Absolute Error) can be captured by the inclusion of competitive effects of 12% with cross-category effects of 0.6%.

The large variety of promotion types and complex promotional interactive effects make product sales difficult to forecast in formal model based approaches. Methods exist to incorporate the large number of promotional related variables (e.g. Ma, et al., 2016) into an operational forecasting model as is necessary for retail promotion optimization, and software solutions exist that implement such models, e.g., the SAP Customer Activity Repository solution for retail (http://help.sap.com/car), which calculates causal forecasts with an elaborate promotion model using regularization. Nevertheless, there is little evidence of widespread use of such models. Instead, many retailers use a simple statistical model supplemented by judgmental adjustments by the demand planning team (Fildes, Goodwin, Lawrence, and

Nikolopoulos, 2009) where one of the companies analyzed was a retailer. An important research issue is what if any benefits accrue from the increasingly complex alternative methods.

Online product reviews and social media

Online product reviews have been found to be an important source of market research information for online retailers in recent years (Floyd, Ling, Alhogail, Cho, and Freling, 2014) and we speculate this applies to all retailers where service is an important component. Since such reviews are a voluntary expression of consumers' experiences and beliefs about the quality of products and services, consumers rely on online product reviews when making their own purchasing decisions (Chen and Xie, 2008; Zhao, Yang, Narayan, and Zhao, 2013). Researchers have found that there is a strong relationship between online word-of-mouth and product sales, but that the impact of word-of-mouth varies with product category (Archak, Ghose, and Ipeirotis, 2010; Chen, Wang, and Xie, 2011; Chevalier and Mayzlin, 2006; Hu, Koh, and Reddy, 2014; Zhu and Zhang, 2010). Thus, incorporating online product reviews, using tools such as text mining and sentiment analysis, may allow online retailers to add a new layer to their existing predictive models and boost predictive accuracy. However, it should be noted that fake reviews and so-called "sock puppetry" are a concern (Zhuang, Cui, and Peng, 2018) and automatically detecting such fake reviews is an active field of study (e.g., Kumar, Venugopal, Qiu, and Kumar (2018).

Product reviews are mainly textual data which cannot be used directly in a sales forecasting model although numerical summaries are common. The basic idea using product reviews as explanatory variables is to identify the extent of the product review polarity, e.g., strongly positive, strongly negative, and neutral. Rating is a simple form of product review, and retailers often adopt a five-star rating mechanism, e.g., Amazon's product rating system. This method is simple and fast. Another way to detect the polarity of a review is to examine the textual content of the review (Ku, Lo, and Chen, 2007). In this manner, a keyword dictionary needs to be established before analyzing the polarity of reviews using text mining and sentiment analysis techniques (Chern, Wei, Shen, and Fan, 2015). A more efficient way of using product reviews is where text is represented as the collection of its words, ignoring grammar and even word order but keeping multiplicity in a bag-of-words model, dealing with the resulting large number

of predictors using, say, random projection (Schneider and Gupta, 2016). But on-line reviews, once measured, do not generate a uniquely important driver variable as they interact with any promotions, and both have proved important in forecasting in an application to Amazon on-line sales of electronic products (Chong, Li, Ngai, Ch'Ng, and Lee, 2016).

Similar challenges arise if we try to improve retail sales forecasts using social media data, such as Facebook, Twitter, Weibo, blogs or similar services. Here, a social media post first needs to be matched to the corresponding product – in the case of online product reviews on a product's page, it is clear which product a review belongs to. Once this step is taken, similar text mining methods can be brought to bear on this topic as in the case of online product reviews. Care must be taken to distinguish forecasts using user-generated social media data from forecasts using social media data that the retailer (or the manufacturer) created in conducting marketing activities on social media (Kumar, Choi, and Greene, 2016) – both cases can be termed "forecasting with social media", and confusion may result. Evangelos, Efthimios, and Konstantinos (2013) and Harald, et al. (2013) offer reviews on forecasting product sales (and other variables of interest) using social media data.

In either case, given the ephemerality of social media and the difficulty in forecasting customer reviews or social media posts themselves, in contrast to a retailer's own marketing and pricing activities, these variables will likely only offer possibilities to improve short-term forecasts, not medium- or long-term ones. However, the evidence of success is extremely limited (Schaer, Kourentzes, and Fildes, 2018)

4.4 Data pooling

After selecting an appropriate aggregation level, another modeling issue is deciding the appropriate extent of pooling across stores and SKUs. Pooling addresses the data availability issue by leveraging analogous sales time series to learn common patterns (Frees and Miller, 2004). Pooling observations across a group of similar items is expected to lead to higher forecasting accuracy, with fewer parameters to be estimated, adapting more rapidly to any structural changes in time series with robustness in the presence of outlier observations (Duncan, Gorr, and Szczypula, 2001). Retail chains are characterized by a multitude (hundreds) of stores

of different formats in diverse geographic regions. Pooling data across SKUs, subcategories and stores increases the size of the training dataset and the observed ranges for the explanatory variables. The marketing mix elasticities and seasonality patterns are usually assumed to be homogeneous in a pool of stores, but the baseline sales are allowed to be heterogeneous for different stores (Ainscough and Aronson, 1999; Baltas, 2005; van Donselaar, Peters, de Jong, and Broekmeulen, 2016). But the decision as to which variables should be assumed homogeneous (or heterogeneous) is still a matter of judgment. The downside of inappropriately assuming homogeneity is that the forecast equations are mis-specified with a resultant bias.

In the famous PromoCast model, Cooper et al. (1999) used a 67-variable cross-sectional pooled regression analysis of SKU-store sales under a variety of promotional conditions with store and chain specific historical performance information. Andrews, Currim, Leeflang, and Lim (2008) found that accommodating store-level heterogeneity does not improve the accuracy of marketing mix elasticities relative to the homogeneous model, and the improvements in fit and forecasting accuracy are also modest. Gür Ali, Sayin, van Woensel, and Fransoo (2009) compared the accuracy of 30 SKU sales prediction methods differing in data richness, technique complexity and model scope using a multi-store, multi-SKU European grocery sales database. They found that pooling observations across stores and subcategories provided better predictions than pooling across either only stores or only subcategories.

Contrasting with the findings by Andrews et al. (2008) and Gür Ali et al. (2009) that store homogeneous models provided better forecasts, Lang, Steiner, Weber, and Wechselberger (2015) found that allowing for heterogeneity in addition to functional flexibility (P-splines instead of linear) could improve the predictive performance of a store sales model considerably: incorporating heterogeneity alone only moderately improved or even decreased predictive validity.

Another benefit of data pooling is that it can be used to forecast demand for new SKU-store combinations not present in the training data. This then can be used to make allocation decisions for both existing and new stores. For example, Gür Ali (2013) proposed a "Driver Moderator" method which can generate short-term forecasts for both existing and new SKUs by pooling information across SKUs and stores. Similarly, Ferreira, Lee, and Simchi-Levi

(2015) present a pricing DSS for an online retailer, Rue La La, which offers extremely limited-time discounts ("flash sales") on designer apparel and accessories. They used the features from the historical data to build a regression model for each product department, pooling data, starting with a SKU such as women's athletic shoes to the top of the hierarchy, the department, such as footwear. The model then predicts demand of future first exposure styles depending on the price and level of discount as well as the SKU characteristics. While the forecast evaluation was cross-sectional, the successful revenue optimization experiment supported the effectiveness of the demand model.

In addition to data pooling, studies have also found that forecasts could be further improved by mining the residuals from many SKUs pooled across subcategories and stores. Based on the PromoCast model proposed by Cooper, et al. (1999), Cooper and Giuffrida (2000) use data mining techniques on the residuals to extract information from many-valued nominal variables, such as the manufacturer or merchandise category. The output of the data mining algorithm is a set of rules that specify what adjustments should be made to the forecast produced by the homogeneous market-response model. This combination means that a more complete array of information could be used to develop tactical planning forecasts. Trusov, Bodapati, and Cooper (2006) further improve the accuracy of the forecasts and interpretability of the recommendation system for promotional forecasts. Gür Ali and Pinar (2016) proposed a two-stage information sharing method. Segment-specific panel regressions with seasonality and marketing variables pool the data first; 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 forecast combines the extrapolated residuals with the forecasts from the first stage which showed out-of-sample accuracy improvements of 15%-30% over a horizon of 1 to 12 months compared with that of the one stage model. Exponential smoothing provided the benchmark where again the gains were substantial (of between 25%-40%).

When lacking sufficient item level data to develop an item-specific model, the research summarized above has shown pooling to be a good way to improve the forecasts. But if there is enough historical data available for a single SKU, the following questions arise: should we

use an individual model for that SKU or a pooled model considering all other SKUs? And how much historical data is enough? So far there is no systematic research to answer these questions.

4.5 Dimensionality reduction in presence of promotions

Any complete specification of the product (SKU) level determinants of store sales has high dimensionality in the explanatory variable space due to cross-item promotional interactions, which pose a big challenge in product demand forecasting. The model may be easily over-fitted or even cannot be estimated in this situation. The high dimensionality thus mainly stems from competing products within the same category (Ma et al., 2016).

One simple solution is to select the most influential subset of items in the same product category as the focal product. For example, the forecasting model named CHAN4CAST which was developed by Divakar et al. (2005) considers only the main competitor‘s promotional variables, i.e., considering only Pepsi’s promotions when forecasting the sales of Coca-Cola. Similarly, Lang et al. (2015) select the lowest price of a competing national (premium) brand in store as representative of the competition. Ma et al. (2016) used the promotional information from the top five sales products in the same category as the focal SKU, achieving a 6.7% improvement in forecast accuracy compared to models that only used information on the focal variable.

Another way of overcoming the dimensionality problem is to build predictive models based on summaries of the cross-promotional effects. The basic idea of this method is to build indexes which could summarize the cross promotional information. For example, van Donselaar et al. (2016) simply use the number of SKU in promotion as the summary of the promotional intensity in the category. Another straightforward way of building promotional indexes is to construct a weighted averaging of the promotion values (discount, display and feature) across SKUs in the category (e.g., Natter, Reutterer, Mild, and Taudes, 2007). Voleti, Kopalle, and Ghosh (2015) proposed a more elaborate approach by simultaneously incorporating branding hierarchy effects and inter-product similarity.

The third way is to summarize the promotional information by extracting a few diffusion “factors” by Principal Component Analysis (PCA) (Stock and Watson, 2002). A criticism of

factor augmented regressions is that the factors are estimated without taking into account the dependent variable. Thus, when only a few factors are retained to represent the variations of the whole explanatory variable space, they might embody only limited predictive power for the dependent variable whereas the discarded factors might be useful.

Another solution is based on variable selection, especially by penalized likelihood method to automatically select influential promotion variables via continuous shrinkage (Gür Ali, 2013; Huang, et al., 2014; Ma, et al., 2016). Traditional best subset selection procedures are usually infeasible for high-dimensional data analysis because of the expensive computational cost. Penalized likelihood methods have been successfully developed over the last decades to cope with high dimensionality (Friedman, 2012; Tibshirani, 2011). A number of recent studies on product demand forecasting are all based on this method. For example, Gür Ali (2013) proposed a "Driver Moderator" method which uses basic SKU-store information and historical sales and promotion data to generate many features, and an L1-norm regularized regression simultaneously selects a few relevant features and estimates their parameters. Similarly, Huang et al. (2014) also identify the most relevant explanatory variables using L1-norm regularized methods. Ma et al. (2016) proposed a four step methodological framework which consists of the identification of potentially influential categories, the building of the explanatory variable space, variable selection and model estimation by a multistage LASSO (Least Absolute Shrinkage and Selection Operator) regression, followed by a scheme to generate forecasts. The success of this method for dealing with high dimensionality is demonstrated by substantial improvements in forecasting accuracy compared to alternative methods of simplifying the variable space. The multi-stage procedure overcomes the known limitation of LASSO for dealing with highly correlated explanatory variables.

While the issue of dimensionality reduction of the explanatory variable space may seem arcane, simple methods such as stepwise regression do not work and the effects on both forecast accuracy but also overall profitability (through a retail price-promotion optimization) are substantial (Kunz and Crone, 2015). Practical research is needed in dealing with the typical messy data encountered in retail operations where time series histories are short, multiple promotion types are thought to be relevant and product assortments change. There are

potentially major benefits to be gained from estimating the loss from using simplified methods.

5. Product level demand forecasting methods

Much effort has been devoted over the past several decades to the development and improvement of demand forecasting models in retail. Beyond well-established univariate extrapolative methods such as exponential smoothing, linear regression models (and variants) that include various driver variables are preferred over more complex models. Such linear models have the important practical advantage of easy interpretation and implementation. On the other hand, if linear models fail to perform well in both in-sample fitting and out-of-sample forecasting, more complex nonlinear models should be considered (Chu and Zhang, 2003). Indeed they are embedded in some commercial software (Fildes, Schaer, and Svetunkov, 2018). We here review these different classes of such models.

5.1 Univariate forecasting methods

The basic product level demand forecasting methods are univariate forecasting models using only the past sales history. The techniques used in retail product demand forecasting range from traditional time series techniques, such as simpler moving averages, the exponential smoothing family or the more complicated Box–Jenkins ARIMA approach (Kalaoglu, et al., 2015), Fourier analysis (Fumi, Pepe, Scarabotti, and Schiraldi, 2013), to state space models (Patrícia Ramos, Santos, and Rebelo, 2015). As the outputs from SKU level forecasts are used in inventory rules, Taylor (2007) developed an exponentially weighted quantile regression method, which generates interval forecasts from quantile predictions. However, these methods do not take external factors such as price changes and promotions into account. Gür Ali et al. (2009) found that simple time series techniques perform well for periods without focal product promotions. However, for periods with promotions, models with more inputs improve accuracy substantially. Therefore, univariate forecasting methods are usually adopted for higher aggregation demand forecasting (e.g., product category demand at a chain), or for products with low promotion or price elasticity of demand.

5.2 Base-times-lift and judgmental adjustments

In practice, many retailers use a base-times-lift approach to forecast product demand (Cooper et al., 1999). The approach is usually a two-step procedure which initially generates a baseline forecast from a simple time series model and then makes adjustments for any upcoming promotional events. The adjustments are often estimated based on the lift effect of the most recent price reduction and/or promotion, and also the judgements made by retail managers (Cooper et al., 1999). In part, we conjecture, this is due to the installed base of commercial software rather than any appraisal of the effectiveness of more complex algorithms.

The use of analogous past promotions which are regarded as similar to the forthcoming promotion is a natural basis for the forecast (Fildes and Goodwin, 2007; Lee, Goodwin, Fildes, Nikolopoulos, and Lawrence, 2007). McIntyre, Achabal, and Miller (1993) proposed a Case-Based Reasoning system to facilitate promotional adjustments. The system selects the historical analogs that are most similar to the planned promotion, and then adjusts the sales of each analog to account for any differences between the analog and the planned promotion. The forecasts are derived from the multiple analogs to arrive at a single sales projection. Lee et al. (2007) conducted an experiment which showed that a forecasting support system (FSS) could be designed to provide users with guidance on appropriate similarity judgments based on evidence from past promotions: the consequential adjustments led to more accurate forecasts of the effects of sales promotions. But Fildes, Goodwin, and Önkal (2018) show in experiments that such base lift adjustments fail to take into account the full history of similar product promotions.

Judgmental adjustments are common in practice, but expensive at SKU level demand forecasting where the number of adjustments is large and potentially prone to systematic biases and inefficiencies (Fildes and Goodwin, 2007; Fildes et al., 2009). There is a substantial body of literature on the creation and evaluation of judgmental forecasts, but few have been conducted in a retailing context (Lawrence, Goodwin, O'Connor, and Önkal, 2006) though one of the four companies studied by Fildes et al. (2009) was a non-food retail chain who adjusted some 20% of their forecasts. Overall these adjusted retail forecasts were biased and inefficient, with no value added by the adjustment process: a possible explanation is that despite accuracy

being the stated objective many of the adjustments were motivated by stocking/service level considerations. Some studies have shown that judgmental adjustments can enhance baseline forecasts during promotions, but not systematically: more advanced statistical models that include promotional indicators have proved better than the expert adjustments (Lim and O'Connor, 1996; Trapero, et al., 2014; Trapero, Pedregal, Fildes, and Kourentzes, 2013) but the evidence is from manufacturing. Judgmental adjustments are often applied in retailing as we discuss in Section 7, nevertheless the scale of retail forecasting necessitates a more selective approach when considering adjustments than that seen in manufacturing (typically <20% compared to around 70%). This naturally motivates the search for modeling methods that include promotional and other variables. No evidence has been collected from the field as to how the inclusion of such explanatory variables affects the adjustment process, but case vignette 1 in Section 7 demonstrates that the use of an econometric modelling approach does not preclude subsequent adjustment. Experimental evidence suggest adjustments can take into account causal information though they are (as expected) smaller than optimal (Lim and O'Connor, 1996; Sroginis, Fildes, and Kourentzes, 2018). The research issue here is whether software can be designed to ensure expert information is incorporated into the forecast (through demand planning meetings, for example), avoiding double counting and excluding the irrelevant cues which commonly are part of the forecasting support system (FSS) and the associated organizational process.

5.3 Econometric methods

Another stream of studies uses a model-based system to forecast product sales by directly taking into account promotional (and other) information. These methods are usually based on multiple linear regression models or more complex econometric models whose exogenous inputs correspond to seasonality, calendar events, weather conditions, price, and promotion features.

The merit of linear regression is that it is simple, easy and fast to fit, so it is feasible for large scale product level forecasting problems. A variety of forecasting solutions have been

based on regression models with different specifications: multiplicative (log-log), exponential (semi-log) and log-reciprocal functional forms are the most widely used parametric specifications to represent nonlinearities in sales response to promotional instruments. A well-known example is the SCAN*PRO model and its extensions which decompose sales for a brand into own- and cross -brand effects of price, feature advertising, aisle displays, week effects, and store effects (Andrews, et al., 2008; Foekens, et al., 1994; Van Heerde, Leeflang, and Wittink, 2000, 2001). PromoCast is another well-known promotion-event forecasting model which was developed by Cooper et al. (1999). They used a static cross-sectional regression analysis of SKU-store sales under a variety of promotion conditions, with store and chain specific historical performance information. Divakar et al. (2005) also employed a dynamic regression model capturing the effects of such variables as past sales, trend, own and competitor prices and promotional variables, and seasonality.

In addition to linear regressions, more sophistical models that include complex error correlation structures have been proposed. Curry, Divakar, Mathur, and Whiteman (1995) proposed a Bayesian VAR model to forecast canned soup product sales at the brand level. The model included the sales, price, and advertisement of four competing brands as endogenous variables. Baltas (2005) proposed a panel regression model that admits store heterogeneity, periodic sales variation, chain-wide sales shocks, and sales dynamics. Recently, using the IRI data set (Bronnenberg et al., 2008), Huang et al (2014) and Ma et al. (2016) have developed Autoregressive Distributed Lag (ADL) models and evaluated them on SKU data for many categories and a number of stores with the latter study showing that a Lasso procedure could successfully take into account both intra-category promotional variables (12% improvement in MAE) and inter-category (worth a further 0.6%). Arunraj and Ahrens (2015) developed a seasonal autoregressive integrated moving average with external variables (SARIMAX) model to forecast the daily sales of bananas in a German retail store. Michis (2015) proposed a wavelet smoothing method to improve conditional forecasts generated from linear regression sales response models.

With such a variety of models, the key question is what is known about their relative performance: this we consider in Section 5.5 and Table A1.

5.4 Nonlinear and machine learning methods

Nonlinear methods include traditional nonlinear regressions, non- or semi-parametric regressions, and fuzzy and machine learning algorithms. Compared to the linear regression models, nonlinear methods allow arbitrary non-linear approximation functions derived (learned) directly from the data and this increased generality improves the potential to provide more accurate forecasts (though with an increased danger of overfitting).

For grocery products, most published research found improvements in forecasting accuracy by using nonlinear models over linear regressions. The models used include Back Propagation Neural Networks (Aburto and Weber, 2007; Ainscough and Aronson, 1999), Fuzzy Neural Networks (Kuo, 2001); Regression Trees (Gür Ali et al., 2009), Gray relation analysis and multilayer functional link networks (Chen and Ou, 2009, 2011), a two level switching model selecting between a simple moving average and a non-linear predictor (e.g., k-nearest neighbor, decision trees) based on the characteristics of the time series (Žliobaitė, Bakker, and Pechenizkiy, 2012), Support Vector Machines (Gür Ali and Yaman, 2013; Pillo, Latorre, Lucidi, and Procacci, 2016), Wavelets Neural Networks (Veiga, Veiga, Puchalski, Coelho, and Tortato, 2016), and Bayesian P-splines (Lang et al., 2015). An exception where non-linearities led to poor performance is van Donselaar et al. (2016) who analyzed the impact of relative price discounts on product sales during a promotion but did not find conclusive evidence for the presence of threshold and/or saturation levels for price discounts for perishable products. Despite the hype, if non-linearities were commonplace and easy to identify we would expect to see more of such models used in practice.

For apparel or fashion products, nonlinear models have been consistently found to provide better forecasts. Choi, Hui, Liu, Ng, and Yu. (2014) argued that this is due to the demand of apparel and fashion products being notoriously highly volatile, and it is therefore difficult to identify the underlining pattern. Hence the well-established and traditional statistical methods fail to make a sound prediction. Among these techniques, neural networks (NN) are probably the more studied. For example, Au, Choi, and Yu (2008) proposed an evolutionary computation approach in searching for the ideal neural network structure for an apparel sales forecasting

system which they found to be more accurate than a traditional SARIMA model. But the standard gradient learning algorithm estimation such as Back Propagation (BP) NN is relatively more time-consuming. Extreme learning machines (ELM), which provide much faster learning speed, have been adopted in a number of fashion forecasting studies (Wong and Guo, 2010; Xia, Zhang, Weng, and Ye, 2012; Yu, Choi, and Hui, 2011). The experimental results have shown that the performance of the ELM is more effective than traditional BPNN models for fashion sales forecasting but their accuracy compared to BPNN is at best moot. When the historical data is limited, Grey model based methods are claimed to have better performance (Choi, et al., 2014; Xia and Wong, 2014), but the results need to be further validated based on a larger sample of series. Du, Leung, and Kwong (2015) proposed a Multi-Objective Optimization-based Neural Network (MOONN) model which they claimed to be superior to several of the above mentioned methods for the short-term replenishment forecasting problem.

The scalability of the nonlinear models is usually poor so dealing with real retail applications with tens of thousands of SKUs in hundreds of stores is impractical with current computational powers. The amount of training time required to build and maintain nonlinear forecasting models becomes a serious concern. The size of the pooled dataset and memory limitations also raise estimation problems (Gür Ali and Yaman, 2013). For these reasons, existing researchers who have tried to test the superiority of nonlinear models at product level usually work at very small scale (i.e., tens of items). We conclude the evidence for non-linearity generally leading to better forecasting accuracy is weak, the positive evidence probably arising from 'publication bias': The studies cited have many limitations and these are summarized in the next section.

5.5 Comparative evaluation

Table 1 in Appendix A summarizes the studies described in the previous sections, focusing on how thoroughly they have been evaluated. We first describe the focus of the study and the range of products which have been analyzed (in columns 2 and 3). There are established guidelines for the rigorous evaluation of forecasting methods (Tashman, 2000) and these are considered in column (4), in particular the comparison of the proposed model with simple

benchmarks (col. (5)). The authors' conclusions are summarized in the final column. The studies are necessarily selective where we have only included those using retail product level data in their empirical tests and their novelty or extensive testing justified discussion in the previous subsections: a few papers which did not clearly report their experimental settings (e.g., horizon, lead time, or validation sets) have been omitted.

As Table 1 makes clear, we face limited evidence: the studies summarized there have little methodologically in common and most suffer from a lack of generalizability: they are applied to too few SKUs and do not include standard forecasting methods using a well-defined out-of-sample testing procedure over a specified forecast horizon. However, there are some studies which consider a wide range of SKUs, sometimes at store level and we focus primarily on those studies which transcend these concerns.

Univariate benchmarks

Taylor (2007) examined daily SKU demand at a single supermarket excluding all slow (intermittent) items, leaving 256 time series. His proposed novel method of exponentially weighted quantile regression outperformed the benchmarks of exponential smoothing and the optimized company procedure. He offers insight into the data, characterizing the series as having high volatility, intra week seasonality with few showing any signs of trend. Ramos et al. (2015) compared state space and ARIMA when forecasting monthly demand for women's footwear for lead times 1 to 12 with little difference in accuracy. But such aggregate forecasts are of little value to the retailer. Ramos and Fildes (2017) extended the analysis to 988 SKUs in 203 categories and considered a wide range of univariate methods, again excluding intermittent data, with TBATS the best performer. Here the characteristics of the data were similar to Taylor (2007).

Whether or not the SKUs are promoted they are typically highly volatile (particularly at store level). Standard univariate methods such as ETS (Error-Trend-Seasonality, i.e., state space formulations of standard Exponential Smoothing methods and their straightforward generalizations) do not perform well (e.g. Ma et al., 2016). Retailers, especially online retailers with large numbers of active SKUs, face the problem of intermittent demand (even at weekly level) and a vast number of series (535K are used in Seeger et al., 2016). This therefore requires

the tailoring of new methods to these data characteristics (Seeger et al, 2016; Flunkert, Salinas, and Gasthaus, 2017). While the methods are shown to outperform a state space method designed for intermittent data (Snyder, Ord, and Beaumont, 2012), they have not been compared to standard intermittent benchmarks with the Flunkert et al. (2017) study eschewing promotional data. Nevertheless, they show considerable promise.

Research into comparative univariate methods needs to concentrate now on the structural characteristics of retail data: general studies such as the M4 competition (Makridakis, Spiliotis, and Assimakopoulos, 2018) tell us little of direct relevance to the retailer.

Multivariate methods

There are many studies that inform the development of multivariate forecasting methods. Key issues are the variables to consider for inclusion (e.g. intra and intercategory promotional variables, Ma et al., 2016; on-line consumer reviews, Schneider and Gupta, 2016), and the level of pooling used to estimate the elasticities where there is a trade-off between sample size and heterogeneity in the model. New methods such as Michis (2015) examining wavelets or Pillo et al. (2016) considering machine learning methods (Support Vector Machines and Neural Networks) have been proposed, which the researchers claim to be successful, but as we noted, the comparisons made are too limited. In contrast, Gur Ali (2009, 2016) considered a number of machine learning methods compared to the base-lift benchmark and evaluated them over many store×SKU combinations. The Huang et al. (2014) and Ma et al. (2016) studies using ADL models compared the ADL forecasts for various lead times and loss functions with simple smoothing models updated for promotional periods using the last promotional uplift: the econometric models performed best for a range of categories. Often detailed promotional data is unavailable and simpler models must be used so Ramos and Fildes (2017) in an evaluation based on 988 SKUs in 208 categories have shown that various multivariate methods, extended by including the exogenous price information, outperform the univariate methods with gains typically above 10%. For promotional periods the gains are typically higher: in fact, even for non-promotional periods (for the focal product) in Ma et al., (2016) the Lasso based ADL models continued to outperform ETS substantially.

These comparisons cover established products and neglect intermittent SKUs and new products (the latter to be considered in the next section). The studies using Amazon's on-line data are valuable here, applying as they do to intermittent data, though the conclusion from the Flunkert et al (2017) study suggesting that promotional data was unnecessary is surely to be treated with caution, when we know little about the promotional features of the data. The second problem is that the question of reliable benchmarks for intermittent data and an appropriate measure of accuracy (or service) makes generalizing to other situations difficult.

Overall, the multivariate studies show substantial accuracy improvements for SKU level forecasts over univariate benchmarks. While recent studies have included a wide range of categories the studies have mostly focused on groceries, neglecting other product groups. Intermittency and the effects of market instability have yet to be fully explored. Wide ranging evidence of the benefits of machine learning algorithms is needed if we are to believe the hype that both researchers and software companies have generated.

6. New product demand forecasting in retail

Forecasting the demand for new products is a more difficult task compared with the forecasting for existing products because of the lack of direct historical product data. In practice, the forecast error (MAPE) for company forecasts of the market for their own new products has been found in specific cases to be more than 50% (e.g. Brown and Hunter, 1967; Kahn, 2002) over a time horizon of two years or more, though the evidence is limited and has little relevance to the specific shorter- term requirements of retail. Despite the complexity of the task and its relatively low accuracy, such an effort is essential as it drives a variety of multifunctional decisions. These would include purchasing, inventory levels, decisions related to logistics, effect on the overall assortment's profitability, and financial expectations for the new product. The literature on this topic is vast and even by the mid-1980s, Assmus (1984) found the number of methods too numerous to include in his review paper. Since then, many new product forecasting methods (and models) have been developed, and a number of recent reviews on this topic have been made from different perspectives though none focus any attention on the retail

new product decision (Chandrasekaran and Tellis, 2007; Goodwin, Meeran, and Dyussekeneva, 2014; Machuca, Sainz, and Costa, 2014; Meade and Islam, 2006). In general, new product sales forecasting methods can be grouped into three broad categories: (i) the judgmental approach, which entails management judgment based in part on past experience; (ii) the market research approach, where survey data is used to forecast customers' purchasing potential; and (iii) the analogical approach, whereby the forecaster assumes the product will behave as "comparable products" have behaved, a comparison which entails the identification of such comparators and which itself is heavily judgmental. These various techniques for new product forecasting are described for example in Ord, Fildes, and Kourentzes (2017). Based on past survey evidence, customer/market research followed by management judgment were found to be the most common methods used (Gartner and Thomas, 1993; Kahn, 2002). Where sales of hi-tech products were being forecast and where the novelty of the product was associated with greater uncertainty in relation to sales, judgment was particularly important (Lynn, Schnaars, and Skov, 1999).

Forecasting new products by retailers has some of the same general characteristics as the generic problem but also poses some specific issues. Where the retailer also acts as designer and underwriter, the forecasting problem faced is the same as at any manufacturer and we do not discuss this further. However, in the case where there is a decision to be made about adding a product to an existing category, there is a requirement to forecast the overall demand across stores (for a decision relevant time horizon), the cumulative purchase-repeat purchase path and the cannibalization effects on other products in the category. The addition of a new product in all likelihood would be expected to grow the category also and in fact to increase the likelihood of a product being accepted into the retailer's portfolio (van Everdingen, Sloot, van Nierop, and Verhoef, 2011). The decision to take on the product then depends on the retailer's demand forecasts (of SKU, category), the consequential profit forecasts and the support offered to the retailer by the manufacturer (which itself influences demand, e.g., Ze and Bell (2003)).

With no direct historical data, the formal approach most often adopted (we speculate) is the use of analogy. The analogical approach is to forecast a new product by leveraging past histories of similar products. There are two variations:

(1) Models can be fitted to the historical data for the analogs and used to produce forecasts for the new product on the assumption that its adoption will follow a similar time-series pattern (Bayus, 1993). In a Bayesian framework, this can be formalized to use the predecessor's history to provide priors for the new product and then update these priors as the new product's sales are observed (cf., Lenk and Rao, 1990).

(2) Alternatively, especially if the new product is replacing another similar product, the history of the previous item can be used, which is typically carried over in the retailer's data base.

If there is no direct substitute, then an approach which recognizes the distinction is to identify a set of features and attributes of the product that are similar to existing products, such as brand, flavor, color, pattern, price, and its target customer segments, and then to group the demand characteristics from those items to forecast the demand for the new product. Ferreira et al. (2015) applied this approach to an on-line fashion retailer; Schneider and Gupta (2016) applied this to on-line Amazon sales of tablet computers; Tanaka (2010) applied it to data on early sales of books and consumer electronics to forecast longer-term 6 month sales; Wright and Stern (2015) used analogs combined with trial data for various consumer products. Unfortunately the evaluation of these models has not typically focused on the retailer's decision requirements - whether to stock and how much to order. The identification of appropriate analogies is also difficult as Goodwin, Dyussekeneva, and Meeran (2013) have shown (though not in a retail context); it may be that with the plethora of examples in retail and fashion the methods are more successful as Thomassey (2014) claims and as Ferreira et al. (2015) illustrate.

In recent years, researchers have argued that web based analytics are useful for predicting the potential performance of a product launch. This opens a new direction for market research based new product forecasting. For example, Schneider and Gupta (2016) used text mining of consumer reviews of new Amazon products and these improved accuracy one-week ahead, Kulkarni, Kannan, and Moe (2012) found that search data indicating pre-launch consumer interest in a product (movies) was useful in forecasting initial takings. But as Schaer et al. (2018) show in their literature review and empirical work (on computer games) the value if any is only short-term e.g. up to a month.

Test markets are sometimes used by retailers before deciding on a national launch. Research has mainly been concerned with the optimal selection of the sample markets (Mostard, Teunter, and Koster, 2011), and test-market based forecasting models. For instance, Fisher and Rajaram (2000) presents a clustering and linear programming based methodology for selecting stores to conduct the test and creating a season forecast for the chain based on test results. In a contrasting approach, Wright and Stern (2015) showed how a simple model of trial results based on analagous products from the first 13 weeks can be used for other new products and applied in the national launch decision.

Fashion products provide a particular new product forecasting problem. Pre-season a forecast of total product sales is needed. Again, analogous products are typically used while more advanced methods can incorporate product attributes as well as their sales. These attributes are used to cluster past products in order to develop a sales profile for a new products (Thomassey and Fiordaliso, 2006). A Bayesian framework can incorporate various data sources prior to product launch and later those predictions are updated as new data becomes available: as an example, see Yelland and Dong (2014). The priors were developed from past products, and once the early sales data on the new product are available, they are used to update the model parameters.

In summary, new product demand forecasts are an important aspect of retail forecasting. Apart from the fashion industry and more recently, in on-line businesses, research has been limited. Standard approaches are claimed to apply, and where formal methods are used, analogies are at the heart. Some limited survey evidence has been collected to examine the methods in use and has again found that judgment influenced by analogies is the favoured approach. Whether or not the effect of a new product introduction into the category is recognized it is unlikely to be formally modelled.

7. Forecast evaluation

As in all forecasting subdisciplines, forecasts in retail need to be evaluated. The standard forecast accuracy metrics are commonly employed; however, the specific challenges in retail imply that some metrics can be misleading or even unusable. This mainly depends on the level

of aggregation. In addition, as elsewhere and independent of the error measure used, forecasters need to keep the relevant time horizon in mind – in one of our vignettes, the error measures were all one-period, yet the decision lead times were longer.

A commonly employed accuracy metric as we see in our retailing vignettes is the Mean Absolute Percentage Error (MAPE), mainly for its ease of interpretation and comparability between series on different scales, which is especially important for practitioners who need to explain a forecast and its accuracy to non-specialists in forecasting. However, the MAPE is undefined if there are zero realizations, so it can only be used one sufficiently highly aggregated time series.

A related measure is the weighted MAPE (wMAPE), which can also be expressed as the ratio between the Mean Absolute Error (MAE) and the mean of the actual observations (Kolassa & Schuetz, 2007). This is again scale-free and can be interpreted as a percentage, and it is defined whenever at least one nonzero realization happens in the evaluation period.

Non-scale-free accuracy measures include the Mean Absolute Error (MAE) and Root MSE (RMSE). Because of their scale dependence, these are mainly used to compare the performance of different forecasts methods on a *single* time series, and so are usually unsuitable for use in retail forecasting, where we usually have multiple or many time series. However, MAE is sometimes used to summarize the sales across similar products weighted for example by pack size or price.

One approach to making the MAE or (R)MSE scale-free is scaling it by an appropriate factor, like the mean of the actuals in the evaluation period (which in the case of the MAE leads to the wMAPE, as noted above), or the overall mean. An alternative is scaling them by the corresponding error of a benchmark forecasting method, leading to *relative* errors with respect to this particular benchmark. Such methods require the comparison to be made on the same out-of-sample data and if measured by a geometric mean are readily interpretable as the percentage one method outperforms the other. (See Ord, et al., ,2017 for further details and references). However, one problem with almost all these accuracy measures has been very much underappreciated. Specifically, since forecasting inherently deals with uncertainty, let us “take a step back” and consider the predictive density that is explicitly or implicitly underlies a

Commented [FR1]: I do not think MASE works or is interpretable!

particular forecasting method. Given this density, each point forecast accuracy measure is minimized in expectation by a specific functional of the density (Gneiting, 2011). For example, the MSE and RMSE are minimized by the expectation, and the MAE is minimized by the median of the predictive density (Hanley et al., 2001). Thus, if the predictive density is asymmetric, minimizing the MAE may lead to biased forecasts. Even for only somewhat intermittent demand, the MAE has frequently been found to be minimized by a flat zero forecast (Morlidge, 2015; Kolassa, 2016). As we note in the discussion of intermittent demand, measures that relate directly to the ordering and distribution decision can be used in such cases. Our vignettes suggest they are seldom applied: instead, standard error measures are used in an ad hoc way, potentially undermining their calculation and their value. Theoretically better is to separate out the forecasting problem from the inventory/ ordering decision by using predictive densities, as yet, an unrealistic aspiration despite their theoretical virtues. Proper scoring rules are then needed for comparing the results (Kolassa, 2016) and these can be applied for non-stationary retail data driven by, for example, promotions.

Thus, when evaluating point forecasts, we should keep in mind that common accuracy measures may appear to be easily interpretable, but may actually be misleading if our goal is an unbiased forecast, especially for very fine granular data. Forecasters should therefore report bias measures along with MAPEs or similar KPIs.

A recent proposal to address these shortcomings of classical point forecast accuracy KPIs especially for intermittent demands is to use rate-based errors (Kourentzes, 2014), which assess whether forecasts for an intermittent series are correct *cumulatively* over increasing time horizons. This is an interesting idea, although the interpretability of such measures is unclear, and there seems to have been little adoption of these measures.

In addition, recall the high importance not only of expectation point forecasts, but also of high quantile forecasts, in particular when forecasts are used for replenishment and to yield safety amounts (see for example, Taylor, 2007). Note also that replenishment cycles may cover multiple forecasting time buckets, so to calculate safety amounts, we need quantiles of cumulative forecasts, or convolutions of predictive densities. In practice, though only anecdotal evidence is available, the formulae embedded in most commercial systems for estimating such

quantiles are incorrect (Prak & Teunter, 2019).

8. Retail forecasting practice

No recent surveys of forecasting practice in retail have been carried out, following Peterson (Peterson, 1993) who found limited use of econometric methods – judgment was the most used followed by simple univariate methods. McCarthy, Davis, Golicic, and Mentzer (2006) reached much the same conclusion in a general survey including retail. A more tangential approach is to examine what software suppliers offer. The largest software providers such as SAP and SAS offer a full range of products in their demand planning suites, starting with simple univariate methods, but more advanced multivariate models are also available in modules that are either additional (at extra cost) to the popular base suite or require tailoring to the company's data base and cannot be used automatically. For example, SAP now offer Customer Activity Repository (CAR) though APO or other software providing only univariate procedures remains common. Specialist providers such as Relex perhaps provide a pointer to how retail practice is expected to change: they add machine learning into the mix of regression based methods and its inclusion is becoming an industry standard (though quite what it adds to the suite of available methods is, as we have noted from the empirical comparisons is far from clear). But changes in practice are typically slow. For instance, collaboration between retailers and suppliers has long been a hot topic for both academics and software providers. Yet Weller and Crone (2012) showed through a rigorous survey how little this innovative sharing of retailers' EPOS data and forecasts has impacted on manufacturing practice and such collaborations can be fraught with difficulties. Nor in the vignettes of practice described below was it seen as a major opportunity.

A second aspect of the diffusion of new modelling practices into industry is the need for trained staff. While packages such as SAP's CAR and methods such as those proposed by Ma et al. (2016) attempt to make modelling automatic, there remains substantial doubt as to how acceptable such forecasts are in organizational practice. Without staff trained in regression-based methods the likelihood is that such methods will not be fully accepted as a base for the decisions that need to be made: judgmental overrides remain common from the operational

demand forecasts to the tactical and strategic (such as site location) where the evidence is that while models are used, expert judgment remains a major component (Wood, 2013). The stylized cases we present below covering demand planning activities show that judgment remains a key feature, even when the model base in the FSS includes many causal drivers (e.g. promotional types, holidays). Sroginis et al. (2018) provides experimental evidence that model based causal forecast are not fully responded to and judgment is still an important feature.

In order to understand more as to how the alternative approaches have been implemented and the challenges faced we describe briefly five archetypal chain-store forecasting organizations, based on detailed discussions: we also use the format to discuss additional issues we have identified in interviews with a further three companies. A summary is presented in Appendix B.

Additional evidence has been provided in presentations by Seaman (2018) relating to Walmart and by Januschowski (2017) for Amazon. Both presenters represent data science specialist support and have therefore focused more on the algorithms: in Amazon the aim is a hands-off approach. But the scale of both operations is a constraining factor on the algorithms that can be adopted (in Walmart, a Dynamic local linear seasonal model, in Amazon, this includes an autoregressive neural net).

These vignettes capturing practice demonstrate a number of key points:

(i) Despite the availability of commercial software including sophisticated causal modeling and non-linear methods (sometimes included with the ill-defined term, 'demand sensing'), uptake of such advanced tools has been uneven. It is not clear whether these tools offer advantages commensurate with the higher total cost of ownership in a real retail forecasting situation and few companies have routinized the use of these more advanced procedures; promotional modelling at SKU or SKU x Store level remains simplistic with limited use of lags, promotional definitions and intercategory data.

(ii) New product forecasting remains heavily judgmental and informal.

(iii) Intermittent demand is a key problem where current research has not been adopted.

(iv) KPIs and accuracy measurement is typically not given sufficient attention with data

that may well be heavily influenced by intermittency and extremes. In most of our case data, MAPE is used which is particularly prone to distortion while WMAPE used to give additional weight to high-value or volume series fails to overcome this problem (see Kolassa, 2016). The use of MAPE or MASE can also lead to incorrect conclusions as to the adequacy of the benchmark compared to its competitor (see Davydenko and Fildes, 2013). Volatility in updated forecasts and consequential orders is potentially important for supply chain planning (Seaman, 2018) but this is seldom part of the appraisal. Lead time issues linked to the supply chain are rarely considered.

(v) The area of demand planning in retailing is manpower intensive where staff may have overly limited technical expertise. In a recent development some of the companies have introduced a separate data science group. There is some variability in the role, where in some cases top down processes have been established whilst others take more of a consulting role. In most retailers, forecasting is still dominated by either IT or business, with very limited statistical or forecasting knowledge."

(vi) Judgmental intervention superimposed on model based forecasts remains a significant element in retail forecasting beyond the requirements of new product forecasting. No examples of bricks-and-mortar retailers were found that relied on the fully-automatic use of software.

(vii) More tentatively (based on our convenience sample of retailers and software suppliers), the diffusion of best practice modelling remains slow, perhaps due to the installed base of legacy software and the rarity of formally trained forecasters, statisticians or data scientists among retailers.

9. Conclusions and Future research

The retail sector is experiencing seismic change. At the strategic level, existing bricks-and-mortar retailers face hard choices with regard to their stores and their embrace of on-line activities. The forecasts, naïve though they are, shown in Figures 1 and 2, underline the

continuing potential for disruption. The research literature has not been engaged with the questions posed by such rapid structural change in the development of on-line competition with most though not all companies, moving from presenting a primarily brick-and-mortar offering possibly supplemented with catalog sales (primarily with a fashion goods focus) to one where on-line has both superseded and expanded on the previous range of catalog offerings. Amazon looms over the whole sector, particularly in the US, where its purchase of Whole Foods and its move into physical stores has generated shockwaves. This has presented a variety of forecasting challenges. At one level, each on-line retailer has a developing time series history of products that can be used with standard forecasting methods. However, the addition of an on-line alternative channel by an established brick-and-mortar retailer raises fresh problems and the attempt to make the customer view the alternatives as seamless ("Omnichannel" shopping) raises the question of complementarity and substitution. The decision to launch an on-line service should be based on a forecast of total sales and profits for the two alternatives over the planning horizon. Anecdotal evidence suggests that the primarily reason, in contrast, may well be strategic: 'everyone is doing it'. No studies have been found which examine this problem, however, Hernant and Rosengren (2017) have studied the effects post-launch on customer behavior, both in-store and on-line, concluding "for existing customers, the interaction between average transaction, purchase frequency, and regularity turned out to be a zero-sum game". More positively, the addition of an on-line service offered the opportunity to gain new customers who had not purchased in-store. The results do not generalize though, since they are known to be category specific (Brynjolfsson and Rahman, 2009; Wang, Song, and Yang, 2013). Critically, the final penetration into each category is unknown and again, the question has not been researched. For a retailer contemplating developing an on-line service a hierarchical approach would be needed. However, anecdotal evidence (including one of our case vignettes) suggests the two channels are forecast separately.

A second question raised by the dramatic changes in retail shopping habits in some countries is the mix of store locations (mega store out-of-town, supermarket and local convenience store). The established location model methods which had high credibility have been undermined by the change in consumer behavior, leaving a data base devoid of directly

relevant data on which new models can be constructed.

While the strategic issues facing retailers have been largely neglected by researchers, the operational questions of identifying more effective forecasting methods have been partially resolved. First and most important, even in non-promoted periods the choice of benchmark is important, in part of course because small improvements over many thousands of SKUs lead to major financial and service benefits. Whilst much earlier research was limited by its breadth, more recently, work has considered a range of stores and many SKUs in different categories, thus helping with generalizations: conclusions rarely apply to all categories whilst we speculate there may be greater robustness across stores and SKUs within category. Further research seeking explanations for differing relative performance would be interesting, linking technical forecasting issues to the characteristics of the market (see for example, Huang, Fildes and Soopramanien, 2018). When promotional events are considered, again research has demonstrated the benefits of more complex modelling approaches (see e.g., Gür Ali and Pinar, 2016; Ma, et al., 2016). Most important, these models can be used across categories for price/promotion revenue optimization (Ma and Fildes, 2017). The value of these more advanced models can be demonstrated by the value-added that comes from the inclusion of additional classes of variable (e.g. model dynamics were worth an additional 3.42% in profit (Ma and Fildes, 2017). They also can be used to understand the importance of business pricing rules where companies seem to still rely on ad-hoc procedures (Watson, Wood, and Fernie, 2015).

There remain operational forecasting challenges for omni-retailers such as how the geographical mix of on-line sales affects the optimal warehousing and distribution network. Some, such as Amazon, face a problem of scale (Januschowski, 2017) but in principle the methods described elsewhere work. Amazon employs a 'deep learning' neural network approach to resolve this problem, claiming a 15% improvement over a base-line state space model with a model that includes seasonality, holidays and price changes.

The fourth major challenge, but also an opportunity that researchers have in fact partially embraced, is the ready availability of 'big data' and its potential for improving demand forecasting: every kind of customer behavior, be it in-store or through on-line activity, can be integrated into SKU demand forecasts (Boone, Ganeshan, Jain, and Sanders, 2018). And some

have claimed that this route is set to undermine more conventional demand forecasting by shifting attention onto the individual customer. But as Kolassa (2017) has pointed out, the technical hurdles of translating individual (as opposed to segmented) demand histories into aggregate SKU level forecasts have yet to be successfully overcome.

A second marketing problem arises with on-line purchasing where the web habits of customers, both actual and potential, can be analyzed and linked to purchasing (Kathyayani and Gonuguntla, 2014). This leads to the question of how web sites should be designed to increase their effectiveness.

Whether the value of this behavioral data is a chimera or a research gap soon to be convincingly filled remains to be seen. The related issue of customer reviews, Google searches and other on-line forums and their impact on sales is reviewed by Schaer et al. (2018) who demonstrate that much of the existing research has transgressed the principles of good forecasting research, failing to focus on out-of-sample comparative evaluations at decision relevant lead times.

On-line purchasing behavior raises specific operational problems for the forecaster which are not relevant in bricks-and-mortar alternatives, in particular returns and fraudulent purchasing behavior, both of which show high prevalence and are category specific. The problems can be analyzed using data mining techniques where time series behavioral data are included.

In our research underpinning this article we consulted with various retailers as to what they see as their major forecasting challenges. Despite the contrast between the methods most commonly used (exponential smoothing with promotional profiles and event dummies) and the likely most accurate models, there was little dissatisfaction with the models in use. However, two issues seemed particularly prominent: methods suited to intermittent demand and second, scalability (see e.g. Seaman, 2018) where the algorithms must be run overnight often for upward of 40K SKUs for many hundreds of stores. As a consequence, computational resources are a limitation. A further issue highlighted by the retail supply chain experts interviewed was the increasing importance of new products with short-life time histories.

For researchers, a number of important problems remain under-researched in addition to

the strategic, discussed earlier. The relationship between the SKU level demand across channels for different categories is important as a practical problem while the search for market based explanations make it an interesting theoretical research area. The second operational question is developing methods to support new product introductions and the implications for the supply chain.

For software developers, the issues that this review raises are how to develop automatic scalable models that are robust to the data limitations common in retail operations. Issues of data hierarchies where research solutions exist have seen limited implementation.

Finally, with vast quantities of data increasingly linked to the 'big data' generated by observed consumer behavior, we expect retail forecasting to provide a test-bed for the integration of micro-data into more aggregate demand forecasts.

References

Aburto, L., & Weber, R. (2007). Improved supply chain management based on hybrid demand forecasts. *Applied Soft Computing, 7*, 136-144.

Agrawal, N., & Smith, S. A. (1996). Estimating negative binomial demand for retail inventory management with unobservable lost sales. *Naval Research Logistics, 43*, 839-861.

Ailawadi, K. L., Harlam, B. A., César, J., & Trounce, D. (2006). Promotion profitability for a retailer: The role of promotion, brand, category, and store characteristics. *Journal of Marketing Research, 43*, 518-535.

Ainscough, T. L., & Aronson, J. E. (1999). An empirical investigation and comparison of neural networks and regression for scanner data analysis. *Journal of Retailing and Consumer Services, 6*, 205-217.

Alexander, A., Cryer, D., & Wood, S. (2008). Location planning in charity retailing. *International Journal of Retail & Distribution Management, 36*, 536-550.

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*, 147-156.

Andrews, R. L., Currim, I. S., Leeflang, P. S. H., & Lim, J. (2008). Estimating the SCAN∗PRO model of store sales: HB, FM or just OLS? *International Journal of Research in Marketing, 25*, 22-33.

Archak, N., Ghose, A., & Ipeirotis, P. G. (2010). Deriving the pricing power of product features by mining consumer reviews. *Management Science, 57*, 1485-1509.

Arunraj, N. S., & Ahrens, D. (2015). A hybrid seasonal autoregressive integrated moving average and quantile regression for daily food sales forecasting. *International Journal of Production Economics, 170*, 321-335.

Assmus, G. (1984). New product forecasting. *Journal of Forecasting, 3*, 121–138.

Au, K. F., Choi, T. M., & Yu, Y. (2008). Fashion retail forecasting by evolutionary neural networks. *International Journal of Production Economics, 114*, 615-630.

Aye, G. C., Balcilar, M., Gupta, R., & Majumdar, A. (2015). Forecasting aggregate retail sales: The case of South Africa. *International Journal of Production Economics, 160*, 66-79.

Baltas, G. (2005). Modelling category demand in retail chains. *Journal of the Operational Research Society, 56*, 1258-1264.

Bayus, B. L. (1993). High-definition television: Assessing demand forecasts for a next generation consumer durable. *Management Science, 39*, 1319-1333.

Bechter, D. M., & Rutner, J. L. (1978). Forecasting with statistical models and a case study of retail sales. *Economic Review, 63*, 3-11.

Benito, Ó. G., Gallego, P. A. M., & Kopalle, P. K. (2004). Asymmetric competition in retail store formats: Evaluating inter- and intra-format spatial effects. *Journal of Retailing, 81*, 59–73.

Beule, M. D., Poel, D. V. D., & Weghe, N. V. D. (2014). An extended Huff-model for robustly benchmarking and predicting retail network performance. *Applied Geography, 46*, 80-89.

Birkin, M., Clarke, G., & Clarke, M. (2010). Refining and operationalizing entropy-maximizing models for business applications. *Geographical Analysis, 42*, 422-445.

Boone, T., Ganeshan, R., Jain, A., & Sanders, N. R. (2018). Forecasting sales in the supply chain: Consumer analytics in the big data era. *International Journal of Forecasting*, in press.

Borin, N., & Farris, P. (1995). A sensitivity analysis of retailer shelf management models. *Journal of Retailing, 71*, 153-171.

Bottani, E., Bertolini, M., Rizzi, A., & Romagnoli, G. (2017). Monitoring on-shelf availability, out-of-stock and product freshness through RFID in the fresh food supply chain. *International Journal of RF Technologies, 8*, 33-55.

Boylan, J. E., Chen, H., Mohammadipour, M., & Syntetos, A. (2014). Formation of seasonal groups and application of seasonal indices. *Journal of the Operational Research Society, 65*, 227-241.

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). The IRI marketing data set. *Marketing Science, 27*, 745-748.

Brown, F., & Hunter, R. C. (1967). The relationship of actual and predicted sales and profits in new-product introductions. *Journal of Business, 40*, 233-233.

Brynjolfsson, E., & Rahman, M. S. (2009). Battle of the retail channels: How product selection and geography drive cross-channel competition. *Management Science, 55*, 1755-1765.

Bucklin, R. E., & Siddarth, S. (1998). Determining segmentation in sales response across consumer purchase behaviors. *Journal of Marketing Research, 35*, 189-197.

Chandrasekaran, D., & Tellis, G. J. (2007). A critical review of marketing research on diffusion of new products. *Review of Marketing Research, 3*, 39-80.

Chen, F. L., & Ou, T. Y. (2009). Gray relation analysis and multilayer functional link network sales forecasting model for perishable food in convenience store. *Expert Systems with Applications, 36*, 7054-7063.

Chen, F. L., & Ou, T. Y. (2011). Sales forecasting system based on gray extreme learning

machine with taguchi method in retail industry. *Expert Systems with Applications, 38*, 1336-1345.

Chen, H., & Boylan, J. E. (2007). Use of individual and group seasonal indices in subaggregate demand forecasting. *Journal of the Operational Research Society, 58*, 1660-1671.

Chen, Y., Wang, Q., & Xie, J. (2011). Online social interactions: A natural experiment on word of mouth versus observational learning. *Journal of Marketing Research, 48*, 238-254.

Chen, Y., & Xie, J. (2008). Online consumer review: Word-of-mouth as a new element of marketing communication mix. *Management Science, 54*, 477-491.

Chern, C. C., Wei, C. P., Shen, F. Y., & Fan, Y. N. (2015). A sales forecasting model for consumer products based on the influence of online word-of-mouth. *Information Systems and E-Business Management, 13*, 445-473.

Chevalier, J. A., & Mayzlin, D. (2006). The effect of word of mouth on sales: Online book reviews. *Journal of Marketing Research, 43*, 345-354.

Chiang, J. (1991). A simultaneous approach to the whether, what and how much to buy questions. *Marketing Science, 10*, 297-315.

Chintagunta, P. K. (1993). Using group seasonal indexes in multi-item short-term forecasting. *Marketing Science, 12*, 184-208.

Choi, T.-M., Hui, C.-L., Liu, N., Ng, S.-F., & Yu, Y. (2014). Fast fashion sales forecasting with limited data and time. *Decision Support Systems, 59*, 84-92.

Chong, A., Li, B., Ngai, E., Ch'Ng, E., & Lee, F. (2016). Predicting online product sales via online reviews, sentiments, and promotion strategies: A big data architecture and neural network approach. *International Journal of Operations & Production Management, 36*, 358-383.

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*, 217-231.

Conlon, C. T., & Mortimer, J. H. (2013). Demand estimation under incomplete product availability. *American Economic Journal Microeconomics, 5*, 1-30.

Cooper, L. G., Baron, P., Levy, W., Swisher, M., & Gogos, P. (1999). Promocast (tm): A new forecasting method for promotion planning. *Marketing Science, 18*, 301-316.

Cooper, L. G., & Giuffrida, G. (2000). Turning datamining into a management science tool: New algorithms and empirical results. *Management Science, 46*, 249-264.

Croston, J. D. (1972). Forecasting and stock control for intermittent demands. *Journal of the Operational Research Society, 23*, 289-303.

Curry, D. J., Divakar, S., Mathur, S. K., & Whiteman, C. H. (1995). BVAR as a category management tool: An illustration and comparison with alternative techniques. *Journal of Forecasting, 14*, 181–199.

Curtis, A. B., Lundholm, R. J., & McVay, S. E. (2014). Forecasting sales: A model and some evidence from the retail industry. *Contemporary Accounting Research, 31*, 581-608.

Davies, R. L. (1973). Evaluation of retail store attributes and sales performance. *European Journal of Marketing, 7*, 89-102.

Davydenko, A., & Fildes, R. (2013). Measuring forecasting accuracy: The case of judgmental adjustments to sku-level demand forecasts. *International Journal of Forecasting, 29*,

510-522.

Dehoratius, N., & Raman, A. (2008). Inventory record inaccuracy: An empirical analysis. *Management Science, 54*, 627-641.

Divakar, S., Ratchford, B. T., & Shankar, V. (2005). Chan4cast: A multichannel, multiregion sales forecasting model and decision support system for consumer packaged goods. *Marketing Science, 24*, 334-350.

Du, W., Leung, S. Y. S., & Kwong, C. K. (2015). A multiobjective optimization-based neural network model for short-term replenishment forecasting in fashion industry. *Neurocomputing, 151*, 342-353.

Dubé, J. P. (2004). Multiple discreteness and product differentiation: Demand for carbonated soft drinks. *Marketing Science, 23*, 66-81.

Duncan, G. T., Gorr, W. L., & Szczypula, J. (2001). *Forecasting analogous time series*. In J. S. Armstrong (Ed.), *Principles of forecasting: A handbook for researchers and practitioners* (pp. 195-213). Norwell, MA: Kluwer.

Evangelos, K., Efthimios, T., & Konstantinos, T. (2013). Understanding the predictive power of social media. *Internet Research, 23*, 544-559.

Ferreira, K. J., Lee, B. H. A., & Simchi-Levi, D. (2016). Analytics for an online retailer: Demand forecasting and price optimization. *Manufacturing & Service Operations Management, 18*, 69-88.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces, 37*, 570-576.

Fildes, R., Goodwin, P., Lawrence, M., & Nikolopoulos, K. (2009). Effective forecasting and judgmental adjustments: An empirical evaluation and strategies for improvement in supply-chain planning. *International Journal of Forecasting, 25*, 3-23.

Fildes, R., Goodwin, P., & Önkal, D. (2018). Use and misuse of information in supply chain forecasting of promotion effects. *International Journal of Forecasting*, in press.

Fildes, R., Schaer, O., & Svetunkov, I. (2018). Software survey: Forecasting 2018. *OR/MS Today, 45*.

Fisher, M., & Rajaram, K. (2000). Accurate retail testing of fashion merchandise: Methodology and application. *Marketing Science, 19*, 266-278.

Floyd, K., Ling, R. F., Alhogail, S., Cho, H. Y., & Freling, T. (2014). How online product reviews affect retail sales: A meta-analysis. *Journal of Retailing, 90*, 217-232.

Flunkert, V., Salinas, D., & Gasthaus, J. (2017). DeepAR: Probabilistic forecasting with autoregressive recurrent networks. *arXiv preprint arXiv:1704.04110*.

Foekens, E. W., Leeflang, P. S. H., & Wittink, D. R. (1994). A comparison and an exploration of the forecasting accuracy of a loglinear model at different levels of aggregation. *International Journal of Forecasting, 10*, 245-261.

Frees, E. W., & Miller, T. W. (2004). Sales forecasting using longitudinal data models. *International Journal of Forecasting, 20*, 99-114.

Friedman, J. H. (2012). Fast sparse regression and classification. *International Journal of Forecasting, 28*, 722-738.

Fumi, A., Pepe, A., Scarabotti, L., & Schiraldi, M. M. (2013). Fourier analysis for demand forecasting in a fashion company. *International Journal of Engineering Business*

Management, 5, 5-30.

Gartner, W. B., & Thomas, R. J. (1993). Factors affecting new product forecasting accuracy in new firms. *Journal of Product Innovation Management, 10*, 35-52.

Geurts, M. D., & Kelly, J. P. (1986). Forecasting retail sales using alternative models. *International Journal of Forecasting, 2*, 261-272.

Gijsbrechts, E., Campo, K., & Goossens, T. (2003). The impact of store flyers on store traffic and store sales: A geo-marketing approach. *Journal of Retailing, 79*, 1-16.

Goodwin, P., Dyussekeneva, K., & Meeran, S. (2013). The use of analogies in forecasting the annual sales of new electronics products' *IMA Journal of Management Mathematics, 24*, 407-422.

Goodwin, P., Meeran, S., & Dyussekeneva, K. (2014). The challenges of pre-launch forecasting of adoption time series for new durable products. *International Journal of Forecasting, 30*, 1082-1097.

Gneiting, T. (2011) Making and evaluating point forecasts. *Journal of the American Statistical Association, 106*, 746-762.

Gupta, S. (1988). Impact of sales promotions on when, what, and how much to buy. *Journal of Marketing Research, 25*, 342-355.

Gür Ali, Ö. (2013). Driver moderator method for retail sales prediction. *International Journal of Information Technology and Decision Making, 12*, 1261-1286.

Gür Ali, Ö., & Pinar, E. (2016). Multi-period-ahead forecasting with residual extrapolation and information sharing — utilizing a multitude of retail series. *International Journal of Forecasting, 32*, 502-517.

Gür Ali, Ö., Sayin, S., van Woensel, T., & Fransoo, J. (2009). SKU demand forecasting in the presence of promotions. *Expert Systems with Applications, 36*, 12340-12348.

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*, 471-480.

Hanley, J. A.; Joseph, L.; Platt, R. W.; Chung, M. K. & Belisle, P. (2001). Visualizing the median as the minimum-deviation location. *The American Statistician, 55* , 150-152

Harald, S., Daniel, G. A., Panagiotis, T. M., Eni, M., Markus, S., & Peter, G. (2013). The power of prediction with social media. *Internet Research, 23*, 187-200.

Hartzel, K. S., & Wood, C. A. (2017). Factors that affect the improvement of demand forecast accuracy through point-of-sale reporting. *European Journal of Operational Research, 260*, 171-182.

Hernandez, T., & Bennison, D. (2000). The art and science of retail location decisions. *International Journal of Retail & Distribution Management, 28*, 357-367.

Hernant, M., & Rosengren, S. (2017). Now what? Evaluating the sales effects of introducing an online store. *Journal of Retailing and Consumer Services, 39*, 305-313.

Hu, N., Koh, N. S., & Reddy, S. K. (2014). Ratings lead you to the product, reviews help you clinch it? The mediating role of online review sentiments on product sales. *Decision Support Systems, 57*, 42-53.

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, 738-748.

Huang, T., Fildes, R., & Soopramanien, D. (2018). Forecasting retailer product sales in the presence of structural breaks, Lancaster University working paper.

Huff, D. L. (1963). A probabilistic analysis of shopping center trade areas. *Land Economics, 39*, 81-90.

Hyndman, R. J., Ahmed, R. A., Athanasopoulos, G., & Shang, H. L. (2011). Optimal combination forecasts for hierarchical time series. *Computational Statistics & Data Analysis, 55*, 2579-2589.

Hyndman, R. J.; Koehler, A. B.; Ord, J. K. & Snyder, R. D. (2008) Forecasting with Exponential Smoothing: The State Space Approach. Springer, 2008

Jain, A., Rudi, N., & Wang, T. (2014). Demand estimation and ordering under censoring: Stock-out timing is (almost) all you need. *Operations Research, 63*, 134-150.

Januschowski, T. (2017). Forecasting at Amazon: Problems, methods and systems. In *ISF 2017*. Cairns, Australia: International Institute of Forecasters.

Jin, Y., Williams, B. D., Tokar, T., & Waller, M. A. (2015). Forecasting with temporally aggregated demand signals in a retail supply chain. *Journal of Business Logistics, 36*, 199-211.

Kahn, K. B. (2002). An exploratory investigation of new product forecasting practices. *Journal of Product Innovation Management, 19*, 133-143.

Kaipia, R., Holmström, J., Småros, J., & Rajala, R. (2017). Information sharing for sales and operations planning: Contextualized solutions and mechanisms. *Journal of Operations Management, 52*, 15-29.

Kalaoglu, O. I., Akyuz, E. S., Ecemis, S., Eryuruk, S. H., Sumen, H., & Kalaoglu, F. (2015). Retail demand forecasting in clothing industry. *Tekstil Ve Konfeksiyon, 25*, 174-180.

Kaplan, E. L., & Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457-481.

Kathyayani, C. J., & Gonuguntla. (2014). Cross-format shopping motives and shopper typologies for grocery shopping: A multivariate approach. *International Review of Retail Distribution & Consumer Research, 24*, 79-115.

Kesavan, S., Gaur, V., & Raman, A. (2010). Do inventory and gross margin data improve sales forecasts for us public retailers? *Management Science, 56*, 1519-1533.

Kök, A. G., & Fisher, M. L. (2007). Demand estimation and assortment optimization under substitution: Methodology and application. *Operations Research, 55*, 1001-1021.

Kök, A. G., Fisher, M. L., & Vaidyanathan, R. (2015). Assortment planning: Review of literature and industry practice. In N. Agrawal & S. A. Smith (Eds.), *Retail Supply Chain Management: Quantitative Models and Empirical Studies* (pp. 175-236). Boston, MA: Springer US.

Kolassa, S. & Schütz, W. (2007) Advantages of the MAD/Mean ratio over the MAPE. *Foresight: The International Journal of Applied Forecasting*, *6* , 40-43

Kolassa, S. (2016). Evaluating predictive count data distributions in retail sales forecasting. *International Journal of Forecasting, 32*, 788-803.

Kolassa, S. (2017). Commentary: Big data or big hype? *Foresight: The International Journal of Applied Forecasting*, 22-23.

Koschat, M. A. (2008). Store inventory can affect demand: Empirical evidence from magazine retailing. *Journal of Retailing, 84*, 165-179.

Kostenko, A. V., & Hyndman, R. J. (2006). A note on the categorization of demand patterns. *Journal of the Operational Research Society, 57*, 1256-1257.

Kourentzes, N. (2014) On intermittent demand model optimisation and selection. I*nternational Journal of Production Economics, 156*, 180-190

Kourentzes, N., Petropoulos, F., & Trapero, J. R. (2014). Improving forecasting by estimating time series structural components across multiple frequencies. *International Journal of Forecasting, 30*, 291-302.

Ku, L. W., Lo, Y. S., & Chen, H. H. (2007). Test collection selection and gold standard generation for a multiply-annotated opinion corpus. In *ACL 2007, Proceedings of the Meeting of the Association for Computational Linguistics, Prague, Czech Republic*.

Kulkarni, G., Kannan, P. K., & Moe, W. (2012). Using online search data to forecast new product sales. *Decision Support Systems, 52*, 604-611.

Kumar, N., Venugopal, D., Qiu, L., & Kumar, S. (2018). Detecting review manipulation on online platforms with hierarchical supervised learning. *Journal of Management Information Systems, 35*, 350-380.

Kumar, V., Choi, J. W. B., & Greene, M. (2016). Synergistic effects of social media and traditional marketing on brand sales: Capturing the time-varying effects. *Journal of the Academy of Marketing Science, 45*, 1-21.

Kumar, V., & Leone, R. P. (1988). Measuring the effect of retail store promotions on brand and store substitution. *Journal of Marketing Research, 25*, 178-185.

Kunz, T. P., & Crone, S. F. (2015). The impact of practitioner business rules on the optimality of a static retail revenue management system. *Journal of Revenue and Pricing Management, 14*, 198-210.

Kuo, R. J. (2001). Sales forecasting system based on fuzzy neural network with initial weights generated by genetic algorithm. *European Journal of Operational Research, 129*, 496-517.

Kuvulmaz, J., Usanmaz, S., & Engin, S. N. (2005). Time-series forecasting by means of linear and nonlinear models. In A. Gelbukh, A. DeAlbornoz & H. TerashimaMarin (Eds.), *MICAI 2005: Advances in Artificial Intelligence* (Vol. 3789, pp. 504-513). Berlin: Springer-Verlag Berlin.

Lam, S., Vandenbosch, M., & Pearce, M. (1998). Retail sales force scheduling based on store traffic forecasting. *Journal of Retailing, 74*, 61-88.

Lang, S., Steiner, W. J., Weber, A., & Wechselberger, P. (2015). Accommodating heterogeneity and nonlinearity in price effects for predicting brand sales and profits. *European Journal of Operational Research, 246*, 232-241.

Lawrence, M., Goodwin, P., O'Connor, M., & Önkal, D. (2006). Judgmental forecasting: A review of progress over the last 25 years. *International Journal of Forecasting, 22*, 493-518.

Lee, W. Y., Goodwin, P., Fildes, R., Nikolopoulos, K., & Lawrence, M. (2007). Providing support for the use of analogies in demand forecasting tasks. *International Journal of Forecasting, 23*, 377-390.

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

Levén, E., & Segerstedt, A. (2004). Inventory control with a modified croston procedure and erlang distribution. *International Journal of Production Economics, 90*, 361-367.

Levy, M., Weitz, B. A., & Grewal, D. (2012). *Retailing Management* (8th Edition ed.). New York: Irwin.

Li, C., & Lim, A. (2018). A greedy aggregation–decomposition method for intermittent demand forecasting in fashion retailing. *European Journal of Operational Research*, 269 860-869.

Li, Y., & Liu, L. (2012). Assessing the impact of retail location on store performance: A comparison of wal-mart and kmart stores in cincinnati. *Applied Geography, 32*, 591-600.

Lim, J. S., & O'Connor, M. (1996). Judgmental forecasting with time series and causal information. *International Journal of Forecasting, 12*, 139-153.

Lv, H. R., Bai, X. X., Yin, W. J., & Dong, J. (2008). Simulation based sales forecasting on retail small stores. In *Winter Simulation Conference, Global Gateway To Discovery, WSC 2008, Intercontinental Hotel, Miami, Florida, USA, December* (pp. 1711-1716).

Lynn, G. S., Schnaars, S. P., & Skov, R. B. (1999). Survey of new product forecasting practices in industrial high technology and low technology businesses. *Industrial Marketing Management, 28*, 565-571.

Ma, S., & Fildes, R. (2017). A retail store sku promotions optimization model for category multi-period profit maximization. *European Journal of Operational Research, 260*, 680-692

Ma, S., & Fildes, R. (2018). Customer flow forecasting with third-party mobile payment data. Lancaster University Working paper.

Ma, S., Fildes, R., & Huang, T. (2016). Demand forecasting with high dimensional data: The case of sku retail sales forecasting with intra- and inter-category promotional information. *European Journal of Operational Research, 249*, 245-257.

Machuca, M. M., Sainz, M., & Costa, C. M. (2014). A review of forecasting models for new products. *Intangible Capital, 10*, 1-25.

Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, in press.

McCarthy, T. M., Davis, D. F., Golicic, S. L., & Mentzer, J. T. (2006). The evolution of sales forecasting management: A 20-year longitudinal study of forecasting practices. *Journal of Forecasting, 25*, 303-324.

McIntyre, S. H., Achabal, D. D., & Miller, C. M. (1993). Applying case-based reasoning to forecasting retail sales. *Journal of Retailing, 69*, 372-398.

Meade, N., & Islam, T. (2006). Modelling and forecasting the diffusion of innovation – a 25-year review. *International Journal of Forecasting, 22*, 519-545.

Merino, M., & Ramirez-Nafarrate, A. (2016). Estimation of retail sales under competitive location in mexico. *Journal of Business Research, 69*, 445-451.

Michis, A. A. (2015). A wavelet smoothing method to improve conditional sales forecasting. *Journal of the Operational Research Society, 66*, 832-844.

Moriarty, M. M. (1985). Retail promotional effects on intra-and interbrand sales performance. *Journal of Retailing, 61*, 27-47.

Morlidge, S.(2015) Measuring the quality of intermittent demand forecasts: It's worse than we've thought! *Foresight: The International Journal of Applied Forecasting*, 37, 2015.

Morphet, C. S. (1991). Applying multiple regression analysis to the forecasting of grocery store sales: An application and critical appraisal. *International Review of Retail Distribution & Consumer Research, 1*, 329-351.

Mostard, J., Teunter, R., & Koster, R. D. (2011). Forecasting demand for single-period products: A case study in the apparel industry. *European Journal of Operational Research, 211*, 139-147.

Mulhern, F. J., & Leone, R. P. (1991). Implicit price bundling of retail products: A multi-product approach to maximizing store profitability. *Journal of Marketing, 55*, 63-76.

Murray, K. B., & Muro, F. D. (2010). The effect of weather on consumer spending. *Journal of Retailing & Consumer Services, 17*, 512-520.

Nahmias, S. (1994). Demand estimation in lost sales inventory systems. *Naval Research Logistics, 41*, 739–757.

Natter, M., Reutterer, T., Mild, A., & Taudes, A. (2007). An assortmentwide decision-support system for dynamic pricing and promotion planning in diy retailing. *Marketing Science, 26*, 576-583.

Newing, A., Clarke, G. P., & Clarke, M. (2014). Developing and applying a disaggregated retail location model with extended retail demand estimations. *Geographical Analysis, 47*, 219–239.

Nikolopoulos, K., & Fildes, R. (2013). Adjusting supply chain forecasts for short-term temperature estimates: A case study in a brewing company. *IMA Journal of Management Mathematics, 24*, 79-88.

Orcutt, G. H., & Edwards, J. B. (2010). Data aggregation and information loss. *American Economic Review, 58*, 773-787.

Ord, K., Fildes, R. A., & Kourentzes, N. (2017). *Principles of Business Forecasting* (2nd ed.). New York: Wessex Press Publishing Co.

Osadchiy, N., Gaur, V., & Seshadri, S. (2013). Sales forecasting with financial indicators and experts' input. *Production and Operations Management, 22*, 1056-1076.

Peterson, R. T. (1993). Forecasting practices in retail industry. *The Journal of Business Forecasting, 12*, 11.

Picone, G. A., Ridley, D. B., & Zandbergen, P. A. (2009). Distance decreases with differentiation: Strategic agglomeration by retailers. *International Journal of Industrial Organization, 27*, 463-473.

Prak, D., & Teunter, R. (2019). A general method for addressing forecasting uncertainty in inventory models. *International Journal of Forecasting*, in press.

Pillo, G. D., Latorre, V., Lucidi, S., & Procacci, E. (2016). An application of support vector machines to sales forecasting under promotions. *4OR Quarterly Journal of the Belgian French & Italian Operations Research Societies*, 14, 309-325.

Ramos, P., & Fildes, R. (2017). Characterizing retail demand with promotional effects. In

International Symposium on Forecasting. Cairns, Australia: International Institute of Forecasters.

Ramos, P., & Fildes, R. (2018). An evaluation of retail forecasting methods for promotions. In. Lancaster University Dept. Management Science Working Paper.

Ramos, P., Santos, N., & Rebelo, R. (2015). Performance of state space and arima models for consumer retail sales forecasting. *Robotics and Computer-Integrated Manufacturing, 34*, 151-163.

Reynolds, J., & Wood, S. (2010). Location decision making in retail firms: Evolution and challenge. *International Journal of Retail and Distribution Management, 38*, 828-845.

Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing, 76*, 367-392.

Schaer, O., Kourentzes, N., & Fildes, R. (2018). Demand forecasting with user-generated online information. *International Journal of Forecasting*, in press.

Schmidt, J. R. (1979). Forecasting state retail sales - econometric vs time-series models. *Annals of Regional Science, 13*, 91-101.

Schneider, M. J., & Gupta, S. (2016). Forecasting sales of new and existing products using consumer reviews: A random projections approach. *International Journal of Forecasting, 32*, 243-256.

Seaman, B. (2018). Considerations of a retail forecasting practitioner. *International Journal of Forecasting*, 34, 822-829.

Seeger, M. W., Salinas, D., & Flunkert, V. (2016). Bayesian intermittent demand forecasting for large inventories. In *Advances in Neural Information Processing Systems* (pp. 4646-4654).

Shenstone, L. & Hyndman, R. J. (2005) Stochastic models underlying Croston's method for intermittent demand forecasting. *Journal of Forecasting, 24* , 389-402

Simkin, L. P. (1989). SLAM: Store location assessment model—theory and practice. *Omega, 17*, 53-58.

Smaros, J. (2007). Forecasting collaboration in the european grocery sector: Observations from a case study. *Journal of Operations Management, 25*, 702-716.

Snyder, R. D., Ord, J. K., & Beaumont, A. (2012). Forecasting the intermittent demand for slow-moving inventories: A modelling approach. *International Journal of Forecasting, 28*, 485-496.

Song, Q. (2015). Lessons learned and challenges encountered in retail sales forecast. *Industrial Engineering & Management Systems An International Journal, 14*, 196-209.

Srinivasan, S. R., Ramakrishnan, S., & Grasman, S. E. (2005). Incorporating cannibalization models into demand forecasting. *Marketing Intelligence & Planning, 23*, 470-485.

Sroginis, A., Fildes, R., & Kourentzes, N. (2018). Interpreting algorithmic and qualitative information when making judgmental forecast adjustments. In *International Symposium on Forecasting*. Boulder, Colorado: International Institute of Forecasters.

Stanley, T. J., & Sewall, M. A. (1976). Image inputs to a probabilistic model: Predicting retail potential. *Journal of Marketing, 40*, 48-53.

Steele, A. T. (1951). Weather's effect on the sales of a department store. *Journal of Marketing, 15*, 436-443.

Stock, J. H., & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association, 97*, 1167-1179.

Stock, J. H., & Watson, M. W. (2010). Combination forecasts of output growth in a seven - country data set. *Journal of Forecasting, 23*, 405-430.

Syntetos, A. A., Babai, Z., Boylan, J. E., Kolassa, S., & Nikolopoulos, K. (2016). Supply chain forecasting: Theory, practice, their gap and the future. *European Journal of Operational Research, 252*, 1-26.

Syntetos, A. A., & Boylan, J. E. (2001). On the bias of intermittent demand estimates. *International Journal of Production Economics, 71*, 457-466.

Syntetos, A. A., & Boylan, J. E. (2005). The accuracy of intermittent demand estimates. *International Journal of Forecasting, 21*, 303-314.

Syntetos, A. A., Boylan, J. E., & Croston, J. D. (2005). On the categorization of demand patterns. *Journal of the Operational Research Society, 56*, 495-503.

Tan, B., & Karabati, S. (2004). Can the desired service level be achieved when the demand and lost sales are unobserved? *IIE Transactions, 36*, 345-358.

Tanaka, K. (2010). A sales forecasting model for new-released and nonlinear sales trend products. *Expert Systems with Applications, 37*, 7387-7393.

Tashman, J. (2000). Out-of-sample tests of forecasting accuracy: An analysis and review. *International Journal of Forecasting, 16*, 437-450.

Taylor, J. W. (2007). Forecasting daily supermarket sales using exponentially weighted quantile regression. *European Journal of Operational Research, 178*, 154-167.

Teller, C., & Reutterer, T. (2008). The evolving concept of retail attractiveness: What makes retail agglomerations attractive when customers shop at them? *Journal of Retailing & Consumer Services, 15*, 127-143.

Teunter, R. H., Syntetos, A. A., & Zied Babai, M. (2011). Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research, 214*, 606-615.

Thomassey, S. (2014). Sales forecasting in apparel and fashion industry: A review. In *Intelligent Fashion Forecasting Systems: Models and Applications*. Choi, T-M., Hui, C-L., & Yu, Y. (eds.). Berlin & Heidelberg: Springer-Verlag.

Thomassey, S., & Fiordaliso, A. (2006). A hybrid sales forecasting system based on clustering and decision trees. *Decision Support Systems, 42*, 408-421.

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society, 73*, 273–282.

Ton, Z., & Raman, A. (2010). The effect of product variety and inventory levels on retail store sales: A longitudinal study. *Production & Operations Management, 19*, 546-560.

Trapero, J. R., Kourentzes, N., & Fildes, R. (2012). Impact of information exchange on supplier forecasting performance. *Omega, 40*, 738-747.

Trapero, J. R., Kourentzes, N., & Fildes, R. (2014). On the identification of sales forecasting models in the presence of promotions. *Journal of the Operational Research Society, 66*, 299-307.

Trapero, J. R., Pedregal, D. J., Fildes, R., & Kourentzes, N. (2013). Analysis of judgmental adjustments in the presence of promotions. *International Journal of Forecasting, 29*,

234-243.
Trusov, M., Bodapati, A. V., & Cooper, L. G. (2006). Retailer promotion planning: Improving forecast accuracy and interpretability. *Journal of Interactive Marketing, 20*, 71-81.
van Donselaar, K. H., Peters, J., de Jong, A., & Broekmeulen, R. A. C. M. (2016). Analysis and forecasting of demand during promotions for perishable items. *International Journal of Production Economics, 172*, 65-75.
van Everdingen, Y. M., Sloot, L. M., van Nierop, E., & Verhoef, P. C. (2011). Towards a further understanding of the antecedents of retailer new product adoption. *Journal of Retailing, 87*, 579-597.
Van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2000). The estimation of pre-and postpromotion dips with store-level scanner data. *Journal of Marketing Research, 37*, 383-395.
Van Heerde, H. J., Leeflang, P. S. H., & Wittink, D. R. (2001). Semiparametric analysis to estimate the deal effect curve. *Journal of Marketing Research, 38*, 197-215.
Veiga, C. P. d., Veiga, C. R. P. d., Puchalski, W., Coelho, L. d. S., & Tortato, U. (2016). Demand forecasting based on natural computing approaches applied to the foodstuff retail segment. *Journal of Retailing and Consumer Services, 31*, 174-181.
Voleti, S., Kopalle, P. K., & Ghosh, P. (2015). An interproduct competition model incorporating branding hierarchy and product similarities using store-level data. *Management Science, 61*, 2720-2738.
Vulcano, G., Ryzin, G. v., & Ratliff, R. (2012). Estimating primary demand for substitutable products from sales transaction data. *Operations Research, 60*, 313-334.
Walters, R. G. (1988). Retail promotions and retail store performance: A test of some key hypotheses. *Journal of Retailing, 64*, 153-180.
Walters, R. G. (1991). Assessing the impact of retail price promotions on product substitution, complementary purchase, and interstore sales displacement. *Journal of Marketing, 55*, 17-28.
Wang, Q., Song, P., & Yang, X. (2013). Understanding the substitution effect between online and traditional channels: Evidence from product attributes perspective. *Electronic Markets, 23*, 227-239.
Wang, W. J., & Xu, Q. (2014). A bayesian combination forecasting model for retail supply chain coordination. *Journal of Applied Research and Technology, 12*, 315-324.
Watson, I., Wood, S., & Fernie, J. (2015). "Passivity": A model of grocery retail price decision-making practice. *European Journal of Marketing, 49*, 1040-1066.
Wecker, W. E. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science, 24*, 1043-1054.
Weller, M., & Crone, S. F. (2012). Supply chain forecasting: Best practices & benchmarking study. Lancaster Centre for Marketing Analytics and Forecasting. https://goo.gl/p2vWis.
Weller, M., Crone, S. F., & Fildes, R. (2016). Temporal aggregation and model selection:An empirical evaluation with promotional indicators. In *International Syposium on Forecasting*. Santander, Spain: International Institute of Forecasters.
Williams, B. D., Waller, M. A., Ahire, S., & Ferrier, G. D. (2014). Predicting retailer orders with pos and order data: The inventory balance effect. *European Journal of*

Operational Research, 232, 593-600.

Wong, W. K., & Guo, Z. X. (2010). A hybrid intelligent model for medium-term sales forecasting in fashion retail supply chains using extreme learning machine and harmony search algorithm. *International Journal of Production Economics, 128*, 614-624.

Wood, S., & Browne, S. (2007). Convenience store location planning and forecasting – a practical research agenda. *International Journal of Retail & Distribution Management, 35*, 233-255.

Wood, S., & Reynolds, J. (2013). Knowledge management, organisational learning and memory in uk retail network planning. *The Service Industries Journal, 33*, 150-170.

Wright, M. J., & Stern, P. (2015). Forecasting new product trial with analogous series. *Ssrn Electronic Journal, 68*, 1732-1738.

Xia, M., & Wong, W. K. (2014). A seasonal discrete grey forecasting model for fashion retailing. *Knowledge-Based Systems, 57*, 119-126.

Xia, M., Zhang, Y., Weng, L., & Ye, X. (2012). Fashion retailing forecasting based on extreme learning machine with adaptive metrics of inputs. *Knowledge-Based Systems, 36*, 253-259.

Yelland, P. M., & Dong, X. (2014). Yelland, P. M., & Dong, X. (2014). Forecasting demand for fashion goods: A hierarchical Bayesian approach. In *Intelligent fashion forecasting systems: Models and Applications*. Choi, T-M., Hui, C-L., & Yu, Y. (eds.). Berlin & Heidelberg: Springer-Verlag, (pp. 71-94).

Yu, Y., Choi, T.-M., & Hui, C.-L. (2011). An intelligent fast sales forecasting model for fashion products. *Expert Systems with Applications, 38*, 7373-7379.

Ze, X., & Bell, D. R. (2003). Creating win--win trade promotions: Theory and empirical analysis of scan-back trade deals. *Marketing Science, 22*, 16-39.

Zellner, A., & Tobias, J. (2000). A note on aggregation, disaggregation and forecasting performance. *Journal of Forecasting, 19*, 457-465.

Zhang, G. P., & Qi, M. (2005). Neural network forecasting for seasonal and trend time series. *European Journal of Operational Research, 160*, 501-514.

Zhao, Y., Yang, S., Narayan, V., & Zhao, Y. (2013). Modeling consumer learning from online product reviews. *Marketing Science, 32*, 153-169.

Zhu, F., & Zhang, X. (2010). Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *Journal of Marketing, 74*, 133-148.

Zhuang, M., Cui, G., & Peng, L. (2018). Manufactured opinions: The effect of manipulating online product reviews. *Journal of Business Research, 87*, 24-35.

Žliobaitė, I., Bakker, J., & Pechenizkiy, M. (2012). Beating the baseline prediction in food sales: How intelligent an intelligent predictor is? *Expert Systems with Applications, 39*, 806-815.

Zotteri, G., & Kalchschmidt, M. (2007). A model for selecting the appropriate level of aggregation in forecasting processes. *International Journal of Production Economics, 108*, 74-83.

Zotteri, G., Kalchschmidt, M., & Caniato, F. (2005). The impact of aggregation level on forecasting performance. *International Journal of Production Economics, 93-94*, 479-

491.

Appendix A Product sales forecasting papers

Table 1 A brief summary of representative research papers on product level sales forecasting. The papers, discussed in the text, have been selected as those that contain a wider range of comparisons and more complete evaluations.

Commented [FR2]: How chosen?

OK? Not sure

Reference (1)	Focus (2)	Data: range and granularity in the product, location and time dimensions (3)	Variables (4)	Forecast horizon and evaluation (5)	Baseline methods (6)	Conclusion and notes (7)
Blattberg and George (1991)	Shrinkage estimation of price and promotional elasticities in pooled data. Focus is on "more reasonable coefficient estimates" than on better forecasts	3 grocery chains, 4 brands of bathroom tissue. 110, 112 & 126 weeks of data for the 3 chains	Brand volumes, relative prices to average competitive prices, discount, display, feature, max competitive discount	Final 10%, middle 20% and random 20% of the samples, single origin; relative MSE	Semi-log regression	Pooling brands that have different pricing philosophies yielded more reasonable parameter estimates and improved prediction. Predictive improvement slight for only one pooling alternative. Poor design.
Foekens et al. (1994)	Compare pooling across chains and stores including the homogeneous SCAN*PRO model of retail promotion effects at different levels of aggregation and pooling	3 brands of food products sold in 40 stores from 3 chains, 104 weeks	Sales volume; own and cross prices, own and cross displays & features	52 weeks for estimation, single origin up-to-4, up-to-12 and up-to-52 weeks forecasts; MAPE, MdRAE	Models with heterogeneous response parameters across retail chains	A heterogeneous store model provided superior accuracy at chain and market levels.
Curry et al. (1995)	Forecasting Bayesian Vector Autoregression modeling promotional activity	4 brands in one category (canned soup) on market level (aggregated over 40 stores), 124 weeks	SKU aggregated sales volume: prices, display, feature, advertising exogenous; seasonal indices	Rolling-origin 1-16 week ahead forecasts aggregated in Theil's U	Univariate and multivariate time series models: BVAR, MARMA, ARIMA, Exponential smoothing	BVAR provided more accurate forecasts; MARMA poor, ARIMA competitive.
Cooper et al. (1999)	The implementation of a promotion event forecasting system, PromoCast, and its performance in several pilot applications and validity studies.	Two analyzes. (1) unspecified number of SKUs in in 311 stores, (2) 212 fast-moving SKUs in 178 stores. Time granularity: promotional events spanning one or multiple weeks	70 variables capture effects of promotion style, SKU and store promotion history, and seasonality.	Two longitudinal cross validations with (1) 20,710 and (2) 30,569 promotion events, horizon not reported; evaluation on errors in replenishment cases of typically 12 units	Historical averages; base lift	Regression-style promotion model helped to improve forecasts in large scale applications; store by store forecasts can help retailers with problems of running out of stock or overstocking. Much detail on promotional events.
Ainscough and Aronson (1999).	Examines neural networks as an alternative to traditional statistical methods for the analysis of scanner data	A national brand of strawberry yogurt, 575 weeks of sales data are aggregated from 6 stores	Lagged sales; Price, display, feature, and their two-way interactions.	20% of the sample as the validation set (appox. 20 weeks); MSE	OLS regression	It was found that three- and four-layer neural networks yielded significantly better predictions than OLS regression.

Cooper and Giuffrida (2000)	Examines the value of residual mining on SKU level promotional forecasting	1.6 million grocery SKU records from 96 outlets; focus on the sales forecasting during promotion events that may have various time lengths	Mining with nominal variables, such as manufactures, retailer store, or geographic areas.	A holdout sample of 460,000 records; MAE	PromoCast forecast module	Mining residuals from market response model reduces the promotional forecasts errors.
Dekker et al. (2004)	Using aggregation and combined forecasting to improve seasonal demand forecasts for groups of seasonal products	56 products from three product classes with 5 years of weekly sales data	None	Rolling holdout sample over 52 weeks, lead one week; sMAPE, MAD and MSE	SES; HW	Seasonal indexes that are estimated from the aggregation of similar seasonal time series are helpful in improving the forecasting accuracy.
Trusov et al. (2006)	Evaluates the value of residual mining on SKU level promotional forecasting focusing on promotional adjustment; similar to Cooper & Giuffrida (2000)	8,195 records from the output of the PromoCast Forecast module;	Categorical variables which describe product, store, or promotion conditions.	Forecasting horizon is unclear; 4,000 records as validation correct promotional adjustment.	PromoCast forecast module	Datamining the residuals from response modeling with product attributes improves the forecasts.
Thomassey and Fiordaliso (2006)	Short term sales forecasting for numerous new items	482 textile items with 52 weeks of sales history	Sales histories are used to cluster items with similar patterns; product attributes are used to classify new product to one cluster	285 of 482 used as test items; 1-50 weeks horizon; RMSE, MAPE, MdAPE	Average sales of the items belonging to the same family	Clustering technique and decision tree classifier is useful to forecast sales for new items with limited data.
Taylor (2007)	Forecasting daily supermarket sales using exponentially weighted quantile regression	256 time series of daily sales with median daily sales greater or equal to 5, in length from 72 to 1436 observations	None	Horizon 1 to 14 days; 42,633 post-sample sales observations; Relative MAE, and coverage measures for prediction intervals.	Univariate models (Naïve, Holt's, HW, etc.) and company procedure.	Exponentially Weighted Quantile Regression (EWQR) compared favorably with a variety of other univariate methods. Evaluates robust adaptions of EWQR.
Gür Ali (2009)	SKU demand forecasting in the presence of promotions	168 store-SKU combinations from 4 categories of non-perishable items spanning a period of 76 weeks.	100 features extracted from historical statistics of the past 4-12 weeks to capture category, store, SKU, and promotion characteristics.	Unclear forecasting horizon, 25 weeks of hold-out; only MAE used	Regression & base-lift model	Using more detailed input data is only beneficial if more advanced techniques are used; calls for incorporating SKU data from multiple stores and multiple subcategories into the same model. ES most accurate outside promotions.
Andrews et al. (2008)	Empirical comparisons on SCAN∗PRO models with continuous and discrete representations of heterogeneity when forecasting store-brand sales	5 brands of shampoo from 28 stores, 109 weeks	Price, display, feature and their cross effects	Horizon 1-28 weeks; RMSE	Homogeneous SCAN∗PRO model	Accommodating store-level heterogeneity does not improve the accuracy of marketing mix elasticities relative to the homogeneous SCAN∗PRO model.
Kumar and Patel (2010)	A new method of combining forecasts using clustering. Clusters of items are identified based on the	411 items of women's summer clothing; 21 weeks of data for the selling season	None – price is constant	Rolling from week 5 to week 21, lead 1; average MSE	Individual forecasts based on ES, ARIMA, MA. Results of	Combining forecasts from multiple items using clustering produces better sales forecasts than the individual forecasts.

	similarity in their sales forecasts and then a common forecast is computed for each cluster of items				individual methods not discussed.	
Gür Ali (2013)	Driver moderator method in SKU-store level sales prediction	Two datasets: (i) 451 SKU-store combinations of daily sales over two years; (ii) 1020 SKUs weekly data from IRI dataset over 4 years	Hundreds of features consisting of drivers, their interactions with modifiers, and dummies for months and days.	No horizon is identified: one-fifth of the observations were randomly chosen as validation dataset; MAE & MASE	Neural network, OLS regression, and Regression tree	Driver moderator model with pooled information across SKUs and stores provide more accurate forecasts.
Huang et al. (2014)	The value of competitive information in forecasting FMCG retail product sales	122 SKU aggregated across 83 FMCG stores in 200 weeks	Promotion on focal SKU, Price diffusion index, promotional diffusion index, week dummies, calendar events, sales lags	Rolling out of sample over 70 forecast origins, leads 1,4 & 12; MAE, MASE, SMAPE, MAPE, AvgRelMAE	ADL with/without promotion information, univariate model, & base-lift model	Models integrating competitive explanatory variables with right variable selection process generate substantially more accurate forecasts
Lu (2014)	Selecting appropriate predictor variable and constructing effective forecasting model for computer products	5 computer products, 247 weeks, from a computer product retailer	A variety of variables constructed from sales history, including lags, trends, growth ratios and volatilities.	49 weeks of holdout; single origin; MAD, RMSE, MAPE, RMSPE	SVR, MARS, ARIMA	The hybrid model which uses MARS to select important forecasting variables and then input selected variables to SVR produce better accuracy.
Voleti et al. (2015)	To assess the respective contributions of inter-product competition and brand-SKU hierarchy effects to explaining and predicting demand	96 SKU under 15 brands, beer category data from 56 stores of a midsized grocery chain, 23 weeks	Inter-product competitive effects, marketing mix, month dummies	6 weeks of holdout, single origin, leads 1; RMSE	Standard log-log regression	SKU competition is more local than global in that only subsets of products compete within groups of comparable products
Lang et al. (2015)	Accounting simultaneously for heterogeneity and functional flexibility in store sales models	89 weeks of data on 8 brands in 81 stores from a major supermarket chain	Price, lowest price of a competing national brand, price of a private label brand, display, store-specific random slopes	9-fold cross-validation with 89 weeks data; ARMSE	Parametric/nonparametric regression models with/without considering store heterogeneity	Allowing for heterogeneity in addition to functional flexibility improved the predictive performance of a store sales model
Arunraj and Ahrens (2015)	Interval prediction of perishable food daily sales	5 years of daily sales of banana measured in kilograms from a typical food retail store	Promotion, discount, weather, weekday, month, holidays	Holdout 30% of all the observations; single origin, lead one horizon; MAPE, RMSE, percent of observations fall in 95% prediction intervals	SARIMA, SARIMAX, MLPNN, hybrid SARIMA and MLR	The proposed hybrid SARIMA and Quantile Regression provided better prediction intervals.
Ramos et al. (2015)	Performance comparison between state space and ARIMA models for consumer retail sales forecasting	Monthly sales of 5 categories of women footwear, 64 observations	None	Rolling 1 to 12 months origins, leads 1-12; ME, MAE, RMSE, MPE, MAPE	ARIMA	State space and ARIMA produce similar forecasts

Michis (2015)	Wavelet smoothing method of conditional sales forecasting	one category, one brand, and two SKUs of detergents at national level, 120 weekly observations for each	Weekly and monthly indicators, lagged sales, trend, holidays, average price per unit, total number of promotional packages	30 and 40 weeks as holdout; MSFE, FESD	Conditional regression without smoothing, with the moving-average or Baxter–King filters	Wavelet smoothing provided the best results when applied to highly volatile marketing time series
Chern et al. (2015)	How electronic word-of-mouth affects product sales	52 best-selling cosmetic product，weekly , aggregate over 355 chain stores, 80 weeks	Strongly positive reviews, strongly negative reviews, neutral reviews, sales lag, time	Unclear how the authors set the validation period; MAPE	Moving average forecasts	Products with abundant online reviews obtained better forecasts by proposed online word-of-mouth-based sales forecasting method
Ren et al. (2015)	Fashion products replenishment forecasting	Six fashion items with seven kinds of color, 36 weeks of data from a fashion boutique	Price, sales lags	12 weeks as holdout, leads 1; MSE, sMAPE	ARIMA, RVM, panel regression	Proposed panel data-based particle-filter (PDPF) model shown suitable for conducting fashion sales forecasting with limited data
Du et al. (2015)	Fashion products replenishment forecasting	7 categories of fashion product sold in 3 cities, 17~48 samples for each product, the time dimension is unclear	3 early sales volumes, 1 weather index and 2 economic indices	4~12 samples as holdout for each product; RMSE, MAE, MAPE	Extreme learning machine and two variants	A nondominated sorting adaptive differential evolution algorithm (NSJADE) to optimize the weights of neural networks (NNs) provided superior forecasting performance
Donselaar et al.(2016)	Forecast promotional demand for perishable products	407 SKUs, 86 weeks data from 4 perishable food categories, aggregated on national level from a retail chain	Discount levels, weight of selling unit, shelf life, number of items in promotion, regular price, display, flyer, holiday, baseline sales	Validation on last 24 weeks; single origin; RMSE, MAPE, average bias	Regression models with 5 price discount dummies, linear discount, or quadratic discount. Poor benchmark used	Modeling threshold and saturation effects leads to worse forecasting performance than modeling price reductions linearly or quadratically for perishable products.
Gur Ali and Pinar (2016)	The value of residual extrapolation and information sharing across different stores on store category level forecasting	2330 category-store time series from 336 stores, lengths range from 37 to 60 months	Marketing mixes and calendar effects	Leads 1~12month; Rolling out from 5 origins with the test data consisting of the following 12 months for each origin; MAE, MdAPE, MAPE	ADL, Mixed model with AR(1), & Winter's ES	Residual extrapolation with information sharing across different stores helps to improve the accuracy for category-store forecasting.
Ma et al. (2016)	SKU level retail store sales forecasting considering inter and intra category promotional effects	15 food categories, 926 SKUs from one grocery store, 320 weeks	Sales lags, price, features, displays, calendar events, week dummies	Rolling out of sample over 80 forecast origins, leads 1 & 4 weeks; MAE, RMSE, MASE, MPE, AvgRelMAE	Econometric model only including focus SKU variables.. Univariate ETS model & base-lift model	Inclusion of competitive variables improves forecasting accuracy. LASSO scheme for model simplification is best.
Pillo et al. (2016)	Application of learning machines for sales forecasting under promotions	2 brands of pasta from two different retail stores, 3 years of daily sales	Calendar attributes, promotion, open hours, price, overall number of receipts	Divided last year of sample into 10 test intervals of the same size; rolling for 10 origins; MAPE	Univariate models: ARIMA, ES and HW	Both SVM and ANNs perform better in the case of a suitable inputs selection than the univariate methods.

Commented [KS3]: ?

Seeger et al. (2016)	Large scale intermittent demand forecasting	Two intermittent data bases considered: Automobile spare parts with 20K items, and Amazon SKUs, 40K and 535K	Out-of-stock conditions and calendar effects, price changes	Point predictions over various lead times; quantiles	ETS; Snyder et al. state space model for intermittent demand (Negative binomial)	Generalized linear model designed for intermittent data and including price and promotional effects, seasonality and holidays. Outperformed basic ETS
Veiga et al. (2016)	Applicability of natural computing approaches in foodstuff retail demand forecasting	63 liquid dairy products monthly sales in 3 product group, 108 months, aggregated at national level	None	Horizon 1-12 months;12 months of handout; MAPE and U-Theil	ARIMA and HW	Wavelets Neural Networks provide best forecasts over the benchmarks
Flunkert et al. (2017).	New development of applying Neural Nets to intermittent data in large scale application	Various intermittent data bases considered: together with Amazon SKUs, 40K and 535K: 52 weeks, evaluated on the next year	Product category, seasonality. Price was considered but added little.	Horizons, 1,1- 9. Point predictions over various lead times; quantiles, cumulative demand	Snyder et al. state space model for intermittent demand	AR recurrent NN outperforms baseline and Seeger (2016) approach.
Ramos and Fildes (2017)	Comparative study on univariate and multivariate forecasting methods on storexSKU data	988 SKUs in 203 categories: intermittent demand excluded, 173 weeks	Seasonal data, holidays, promotional events	Horizon 1 week ahead. Rolling origin calculation with updated model parameters. Many error measures included MAPE, RelMAE etc.	ETS, TBATS, ARIMA, Naïve: compared to multivariate alternatives & Lasso	Multivariate model including promotions performed best with TBATS the best of the univariate methods considered.

Abbreviation of forecasting models in the table: ADL：Autoregressive Distributed Lag; ARIMA: AutoRegressive Integrated Moving Average; SARIMAX: Seasonal AutoRegressive Integrated Moving Average with external variables; BVAR: Bayesian Vector AutoRegression; HW : Holt–Winters' procedure ; MARS：Multivariable Adaptive Regression Splines; MLPNN: Multi-Layered Perceptron Neural Network; MLR: Multiple Linear Regression; RVM: Relevance Vector Machine; SARIMA: Seasonal Autoregressive Integrated Moving Average; SES: Simple Exponential Smoothing; SVR：Support Vector Regression.

Abbreviation of accuracy metrics in the table: ME: Mean Error; MAD：Mean Absolute Deviation; MPE: Mean Percentage Error; MdAPE: Median Absolute Percentage Error; RMSE：Root Mean Square Error; ARMSE: Average Root Mean Squared Error; MAPE：Mean Absolute Percentage Error; RMSPE：Root Mean Square Percentage Error; MASE: Mean Absolute Scaled Error; sMAPE: Symmetric Mean Absolute Percentage Error; AvgRelMAE: Average Relative Mean Absolute Error; MSFE: Mean Squared Forecast Error; FESD: Forecast Error Standard Deviation.

Appendix B Retail demand planning vignettes

	Case 1	Case 2	Case 3
Primary focus	Household, FMCG and garden	Food	Fashion
Product range	20K regular products, 20K new or short-season	26K	20K of which 70% regular products. Final forecasts based on SKUs x size profiles.
Data characteristics	2 years data: High intermittency (around 70%), promotional intense (10 types). Some 10% of regular SKUs are promoted weekly for a period of 4-6 weeks.	5 years of data used. 40% intermittent. 45% promoted within the last 40 days.	For regular products, years of history. % intermittent at a store x day level.
Distribution	450+ primarily high-street stores, UK, with 2 distribution centres, plus on-line.	4 distributions centres plus on-line	6 DCs, 350 stores + on-line
Software	Demand forecasts at store level provided by SAP's F&R (Forecasting and Replenishment) module. The F&R module is regression based using dummy variables for the promotional types, seasonality, holidays and other events. *Variation: Aggregated store forecasts used to produce DC forecasts.*	SAS data handling with tailored regression based forecasting module: model takes seasonal (day, month) effects, promotions including shelf space, weather and events. *Variation: AI based software for stores, Relex (extrapolative) software at DC*	JDA Allocation Retail Solutions
Forecasting (established products)	Weekly buckets with 22+ week lead time, pro-rated to daily with forecasts updated daily for retail site ordering. Promotional forecasts 6+ weeks ahead based on agreed promotional plan. *Variation: promotional forecast done 'manually'*	These forecasts take into account the Day of the week, Monthly effect, Promotional effect including additional space, Weather and Events. Forecasts can be overriden by users.	Forecasts weekly (SKU x Store) disaggregated by daily profile. Automatic univariate exponential smoothing based forecast with event dummies -. Promotions forecast outside system.
Forecasting (new products)	Judgment supplemented by analogy	Judgment supplemented by analogy (from similar products sales history). Demand planners will then react as demand is observed to ensure compatability with forecasts.	Judgmental estimates influenced (though not formally) by previous analogous SKUs. Seasonal profile for short-life products.
Forecasting (online sales)	On-line products are forecast separately and are not linked to store sales.	On-line sales merged with local store. *Variation: outsourced. No cannibalization assumed.*	Total SKU sales forecast with a hierarchical forecast by channel.
Inventory data: quality? Used in demand forecasting?			
Demand planning staffing	Approximately 8 FTEs plus 2 analysts with general supply chain planning and forecasting responsibilities. DC forecasting for logistics team. *Varitiona: supported by a data science team*	Category based (with 8 demand planners and 3 supply planners). Central forecasting team of 13 responsible for smooth operations of system forecasts	Central analytical forecasting team of 4? With planning managers in the different business units responsible for the different categories
Collaboration with suppliers	None: *Variation: Order forecast for promotional items shard for delivery to DC. Inventory invofrmation shared with strategic suppliers.*	Information shared on order forecasts based on demand forecasts	
Interventions	F&R is designed to handle multiple promotional types but omits other potentially important factors such as weather and intra-category effects. 5% approximately of system forecasts are adjusted.	Interventions primarily across all stores daily.. In addition individual store adjustments are made leading to 40% of forecasts being affected. *Varition: 99% of forecasts affected*	?% approximately of system forecasts for regular products are adjusted.
KPIs	Accuracy (measured by MAPE by SKU across stores), focused on ? weeks ahead. *Variants: out-of-stock and inventory*	Bias priotiized at store level. *Variation: Sqrt(Sigma(A^2)/Sigma(F^2))*	Out-of-stocks, forecast accuracy, stock level, reach
Key challenges	New product forecasting and short life cycle; weather effects, *variants: promotional effects, intermittent items*	Weather effects. Stock-outs. Identifying changes in dynamic retail market. Unique events. *Variation: hourly forecasts needed*	Daily effects around holidays, monthly effects, new listings, weather, performance/runtime

Notes: WMAPE weights the MAPE across stores of SKUs according to a volume (or value) measure.

Case 4	Case 5
Beauty, body care, healthcare	Convenience Retail, Foodservice
15K regular products; 55M product-locations worldwide	2k regular products, 500 intermittent/seasonal
110 weeks time series	110 weeks of sales history, 5 years of data available in data cubes. 10% of articles are on promotion more than 25% of the year, and an additional 15% are
About 3600 stores and 25 DCs internationally, online	650 stores, 3 Mid-Atlantic DCs; 150 stores, 2 Florida DC's; select forecasting for 3 small DSD companies
SAP F&R for store forecasting and replenishment and DC forecasting, SAP Retail for DC replenishment. Some custom development	SAP F&R (Forecasting and Replenishment) module including promotions, holidays and weather events.
Weekly forecast with daily profiles. No promotions, EDLP strategy	Weekly forecast going out 13 weeks (26 weeks for about 100 stores); weekly forecasts are disaggregated to daily buckets via dayweights.
No forecast in the first week, then new listing algorithms with special safety stock calculation. For relaunches use sales reference logic	New articles are referenced to existing items based on marketing department expectations of sales & patterns.
Use product movements out of the online fulfilment DC as a basis for the weekly forecast	N/A
Good data quality because of QA mechanisms	Quality is intermitent; inventory not taken into account for short shelf life products.
14 FTEs for store forecasting and replenishment	3 FTE Demand Planners, 1 FTE Replenishment Analyst. Also, 1 system admin, 3 associates focused on store level issues including opening/closing new stores and remodels.
Yes	6 week order forecasts provided to external partners. Marketing department provides longer range expectations to key suppliers.
Store and category specific interventions through custom development, e.g., monthly effects, holidays (and surrounding days), sometimes new listings (if conspicuous)	Correctional judgmental adjustments are applied for first time occurrences. Data corrected for weather events. Approximately 75 stores are considered "highly seasonal" and receive extra attention/profile
Accuracy (measured by WMAPE by SKU across stores), Store Order Acceptance, In-stock conditions; do not measure overstocks	Forecast Accuracy - measured by MAPE; Out of Stock occurences and Days of Inventory.
Short shelf life product replenishment. Accurate dayweighting for fresh items. Weather effects. Data Quality. Lack of reporting/root cause analysis.	promotional effects; intermittency; inventory data accuracy & vendor fill rate data.