International Journal of Forecasting 35 (2019) 170–180

Contents lists available at ScienceDirect

International Journal of Forecasting

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

Forecasting sales in the supply chain: Consumer analytics in the big data era

Tonya Boone [a], Ram Ganeshan [a,*], Aditya Jain [b], Nada R. Sanders [c]

[a] Raymond A. Mason Mason School of Business, The College of William and Mary, United States
[b] Zicklin School of Business, City University of New York, United States
[c] D'Amor-McKim School of Business, Northeastern University, United States

ARTICLE INFO

Keywords:
Sales forecasting
Predictive analytics
Literature review
Big data

ABSTRACT

Forecasts have traditionally served as the basis for planning and executing supply chain activities. Forecasts drive supply chain decisions, and they have become critically important due to increasing customer expectations, shortening lead times, and the need to manage scarce resources. Over the last ten years, advances in technology and data collection systems have resulted in the generation of huge volumes of data on a wide variety of topics and at great speed. This paper reviews the impact that this explosion of data is having on product forecasting and how it is improving it. While much of this review will focus on time series data, we will also explore how such data can be used to obtain insights into consumer behavior, and the impact of such data on organizational forecasting.

1. Introduction

Traditionally, forecasts have served as the basis for planning and executing supply chain activities such as sourcing, making, and distributing products and services to customers. Over the years, trends such as the globalization of supply chains, the explosion of product variety, the shortening of product life cycles and increasingly competitive markets have made forecasting more complex, yet its role more critical.

Digital technologies such as advanced Point-of-Sale (POS) systems, the "Internet of Things" (IoT), user-generated content from social media, and cloud computing have enabled firms to collect vast amounts of data in real-time. We use the term "big data" to refer to data sets that are large (large"volume"); collected in near real-time (high "velocity"); and present in a myriad of unstructured forms (great "variety"). The vast amounts of data that are collected and analyzed in near real-time have the potential to improve our understanding of customer behavior, improve demand prediction and better execute supply chain management. The intent of this paper is to provide a review of the way in which this explosion of data is impacting forecasting. For the purposes of this review, we focus on "consumer analytics" from a forecasting perspective, encompassing a set of data-driven techniques that provide insights into consumer buying behaviors. The better a firm understands its customers' buying behaviors, the more accurate its demand forecasts will be, which in turn helps it to plan and execute supply chain operations more efficiently. The context of this review will be that of retailing, and we focus specifically on time series data. However, the ideas described (new data streams, personalization, new variables in forecasting, etc.) are not limited to the customer side of the supply chain, but can also be extended to inter- and intra-company interfaces.

This paper is organized as follows. Section 2 offers an overview of the sources from which vast quantities and new types of data are becoming available. Section 3 discusses existing research and potential future opportunities in consumer analytics, specifically demand forecasting, using the data from these new sources. Section 4 then provides an overview of the challenges of integrating big data

* Corresponding author.
E-mail address: Ram.Ganeshan@mason.wm.edu (R. Ganeshan).

https://doi.org/10.1016/j.ijforecast.2018.09.003
0169-2070/

into the sales and operations planning process, and finally, Section 5 presents a summary and our major conclusions.

2. Sources of big data

For the purpose of exposition, we identify a three-stage path that is followed by a customer when making a purchase. In the first stage, the customer realizes their need for the product and gathers information about the product. In the second stage, the customer interacts with the firm to gather information, and processes it in order to make a purchase decision. Then, in the third and final stage, the customer makes a purchase. In the ensuing discussion, we identify new sources of data that are now available on each of these stages, after which Section 3 reviews the literature and identifies opportunities for research in each of these stages of the customer decision journey.

2.1. Point-of-sale data

Traditionally, POS systems have been used to collect sales data which are then used as a baseline for forecasting future demand. Gone are the days of independent "cash registers" and "credit card" readers: POS systems today provide a so-called "frictionless" checkout experience for the customer and a connected data gathering system for the merchant. Self-checkout and minimally-assisted POS kiosks reduce lines and increase the checkout speed. POS devices are now moving to mobile devices (either NFC-based payments or hardware-assisted such as Square[1]), which is enabling a wider array of retailers to engage with their customers in multiple ways. For example, a customer in an Apple, Inc. retail store can seek help from any retail staff member, and checkout right next to the product using ApplePay, Apple's payment app on a mobile device or a watch. Interestingly, even the POS device is simply a specialized iPhone. In the most extreme case, Amazon Go[2] promises "No Lines. No Checkout" — payment is handled by the Amazon Go App on the customer's phone and sensors detect the contents of the shopping cart. Finally, POS systems are integrated across multiple selling channels, making it possible to blur the differences between the physical and Internet stores.

The advantage of such modern systems is that they now connect the customer (via software applications or through loyalty initiatives) and the firm's customer relationship management (CRM) system directly. Vast amounts of granular data indexed by customer are now readily available.[3] Such data include demographic profiles, order histories, loyalty card information, coupon redemption rates, etc. A second advantage is that modern POS systems are also connected to the inventory and warehouse systems, giving both the retail worker and the customer access to product availability information and the ability to purchase in the broader network of physical and virtual stores. While this may make it difficult to forecast the demand at any particular store, "system-wide" service levels will improve.

2.2. In-store path data

New in-store technologies are blurring the differences between brick-and-mortar and online browsing and purchasing experiences. Data on customers' browsing ("path data") and purchase (intent to buy by adding to cart, abandoning the cart, etc.) behaviors that were once available only to online retailers are now being integrated into physical stores. Beacons that connect to customer smartphones via bluetooth technology identify when a customer has entered the store. For example, Macy's Shopkick App is an example of one such proximity-based engagement. It is triggered by beacons, and when combined with online browsing experience and the customer's position in the store, the store is able to engage customers by pushing content, providing "offers", and awarding prizes.[4] Kroger, one of the largest grocers in the USA, is deploying digital shelf-edges, where personalized information, including pricing, can be triggered as a customer walks by.[5] Retailers such as Marks and Spencer are experimenting with 'Virtual Rails', where the store's entire catalog is displayed electronically on large TV screens. They play related content, and also, as eCommerce customers are accustomed to, suggest additions to the outfit when a customer scans a chosen item. Augmented reality applications are used for virtual dressing rooms,[6] where customers can 'try on' different products and sizes. Traffic counters, infrared sensors and video cameras can now track customer traffic and paths through the store (Section 3.2 explores research in this area). They are used most commonly to reduce checkout lines and schedule and deploy the workforce.

2.3. User-generated content

User-generated content (UGC) commonly refers to data that is created by unpaid contributors. In our context, it can be internet searches such as a Google search, which is available only on an aggregated basis; or it can refer to uploaded Facebook posts, pictures, videos, testimonials, tweets, blog posts, etc., where the identity of the contributor is explicitly available.

UGC has become a critical factor in the product discovery, research, and buy phases of consumer activity. Increasingly, consumers are discovering new products and trends online, trusting and using online reviews and ratings

[1] See squareup.com.

[2] See https://www.amazon.com/b?node=16008589011.

[3] For example, Target, a large US retailer, uses what they call a "Guest ID" to track individual customers. This ID is linked to the customer's browsing and purchase behavior, as well as to their demographic information. See: http://www.nytimes.com/2012/02/19/magazine/shopping-habits.html?pagewanted=1&_r=1&hp. Also see Feng and Shanthikumar (2018) for examples of ways in which such real time data can be used to build forecasting models.

[4] See http://time.com/money/3432693/macys-shopkick-ibeacon/.

[5] See for example https://www.wsj.com/articles/at-kroger-technology-is-changing-the-grocery-store-shopping-experience-1487646362, accessed December 7, 2017. See also the related YouTube video: https://www.youtube.com/watch?v=w2vvcz-fki4, accessed December 7, 2017.

[6] Gap Inc., is experimenting with the DressingRoom by Gap augmented reality-based virtual dressing room. See https://adressed.gapinc.com/blog/gap-ces-announcement-2017-dressingroom-app, accessed July 20, 2017.

of products and services from peers, and engaging more with brands in the presence of UGC.

Firms are taking "digital media" strategies that make use of UGC — search, social media, mobile, and email marketing — seriously. According to one estimate,[7] the digital marketing 'spend' for US companies will be around $120 billion by 2021, representing 46% of all marketing expenditures. Firms are spending significant sums of money to position themselves in search engine searches; and are investing in exploring how best to use social media platforms in conjunction with their most valued consumers (fans, influencers, etc.) to 'spread the word', not only to inform but to persuade potential consumers. There are several case studies in the literature on such efforts that have been successful (for example see Goh, Heng, & Lin, 2013 and Kumar, Bhaskaran, Mirchandani, & Shah, 2013); however, many of these studies do not focus on SKU-level forecasting, creating a need for research in this area.

3. Opportunities for consumer analytics and forecasting

This section discusses existing and potential advances in consumer analytics and demand forecasting that are achieved using the data collected through new sources.

3.1. Point-of-sales data

The data used most often in time series based forecasting are historical data on sales. The obvious attraction of such data is that they are available readily from the most rudimentary of point-of-sales systems. Time series methods are typically used for estimating patterns in past sales data, which are then extrapolated for forecasting future demand. One of the key issues with this method of forecasting is the bias that results from demand censoring: firms do not record information on demand during stock-out, meaning that information on the demand in excess of inventory carried is simply lost. Such demand censoring leads to forecasts which are biased downwards, resulting in poorer inventory decisions, which further increase the extent of censoring. The problem is exacerbated further when the firm is selling multiple products and stock-outs may result in substitutions. In such cases, the firm may end up underestimating the demand for fast-selling products which stock-out often, and overestimating the demand for slow-selling products which do not sell out, but may have higher sales due to substitution.

A number of research papers have highlighted the problem of demand censoring and its adverse effects not only on demand forecasting (Wecker, 1978), but also on the estimation of important consumer characteristics such as price elasticity (Bruno & Vilcassim, 2008), as well as on making important inventory decisions (Tan & Karabati, 2004). Researchers in the area of operations management have proposed a number of advanced econometric methods for correcting this problem, but most have been too complex for practical implementation. In recent years, researchers have proposed various data-driven approaches in order to counter the problem of demand censoring and thus improve forecasting. These approaches are based on the higher level of granularity in sales and inventory data, and new forecasting models that can be applied to such data. Modern Point-of-Sales systems have the ability to capture detailed data on both the timings of sales and the availability of various products in inventory at different points in time. Combining this data allows the firm to see sales as a function of the product availability over time, thus enabling us to obtain a better picture of demand. In a recent paper, Jain, Rudi, and Wang (2015) offered a simple illustration of improvements in forecasting from the use of such granular data. They show that firms can improve their forecasting and inventory decisions significantly by using the data on the timing of sales. They also show that, in many cases, it is sufficient for the firm to capture only the time at which a stock-out takes place. Using the timing of stock-outs allows better estimates of actual sales to be made, improving forecasts. Queenan, Ferguson, Higbie, and Kapoor (2007) offer more practical applications of this idea to the forecasting of demand in the context of revenue management.

This idea of using more granular data on sales and inventory over time has even stronger implications for the development of multi-product demand forecasts. Most firms worry about having both the right assortment and the appropriate inventory for each of the products in the assortment. Solving this problem requires a knowledge of the demands for different products as a function of the assortment offered — effectively, an understanding of customers' inherent demands for different products and their willingness to substitute between them. The availability of more granular data can lead to a great improvement in the way in which sales data can be used for learning about customers' choices between multiple products in an assortment. Using inventory data permits one to deduce the availability of the product assortment over time, and marrying this with the sales transaction data gives a more complete picture of a customer's choice process. The multi-product forecasting method proposed by Karabati, Tan, and Öztürk (2009) uses this idea: in their proposed scheme, the sales data are split into different time intervals, each of which corresponds to the availability to customers of a different product assortment. These disaggregated data are then used to estimate the demand rates for individual products, as well as the substitution probabilities between them. More recent papers have applied this idea of using granular sales data to different types of customer choice models in an attempt to develop a better understanding of the customer choice process; for example, Musalem, Olivares, Bradlow, Terwiesch, and Corsten (2010) consider a multinomial logit (MNL) choice model, van Ryzin and Vulcano (2014) consider a more generalized rank-based choice model, and Chen, Mersereau, and Wang (2017) show how granular sales data can improve pre-launch merchandise testing significantly.

In addition to increasing the granularity of data over time, the modern Point-of-Sales and data storage systems also allow firms to observe the contents of a customer's

[7] See the report from Forrester, a leading research and advisory firm: http://blogs.forrester.com/shar_vanboskirk/17-01-24-us_digital_marketing_spend_will_near_120_billion_by_2021, accessed July 16, 2017.

basket; in other words, which products were bought together and how often. In many contexts (for example grocery retail), customers generally buy multiple products in a single visit, and the availability (or not) of one product may influence the purchase decisions of other products in the customers' baskets. Researchers in marketing have recognized this in their work on the development of models of customer choice (Chung & Rao, 2003; Hruschka, 2017; Russell & Petersen, 2000), but the application of such analyses to the forecasting of product-level demand remains an open avenue for future research.

3.2. In-store data

Next, we consider other sources of in-store data related to customer behavior. These are data related to customers' entry into the store and subsequent interactions with various store elements before making their purchase decision.

3.2.1. Traffic counter data

In recent years, it has become much easier for firms to gather traffic data, i.e., information on how many customers visited their stores at different times. First, many retailers have ventured into electronic retailing, where such data are a lot easier to collect and analyze. Second, electronic traffic counters have become more pervasive and accurate. While many of the applications of traffic data relate to tactical decisions (e.g. store-level staffing decisions), firms are beginning to realize the value of this information for other operational decisions (out-of-store and in-store promotions, salesperson training, store layout design etc.). Traffic data inform the retailer as to how effective their store front and other out-of-store promotions are in attracting customers to step into their store. When combined with sales records, the potential of this data grows many-fold. By combining traffic data with sales data, a firm can also learn about trends in conversion (i.e., what fraction of visiting customers choose to purchase), which in turn can inform it about the effectiveness of different in-store elements in driving conversion. Research papers on this subject (see for example Lam, Vandenbosch, & Pearce, 1998; Mani, Kesavan, & Swaminathan, 2015 and Perdikaki, Kesavan, & Swaminathan, 2012) have shown how sales can be predicted as a function of traffic and store staffing levels. These models can then be used not only for the improved forecasting of aggregate sales (given some information on future traffic), but also for optimizing staffing levels. Another application of traffic data is offered by Lam, Vandenbosch, Hulland, and Pearce (2001), who show the usefulness of such data for evaluating the impacts of different types of promotions.

There are several possible avenues for future research in this area, especially relating to understanding the demand at brick-and-mortar retailer stores. First, while researchers have constructed models that relate sales to traffic, it is not clear how much improvement in accuracy can be achieved by incorporating traffic data. Second, the existing research is also based on highly aggregate (store-level) data, and does little to differentiate between different types of customers entering the store or different types of products being sold. The relationships between these elements and product demand/sales should be explored in future research. Third, the existing research streams are based mainly on the number of customers entering the store, with no attention being paid to either the amount of time customers spend in the store or the number of people in the store. It is well known that the number of people in a store affects the purchase decision – crowds hinder product browsing and access to sales persons, and increase the 'friction' in completing sales transactions. Future research should consider these additional operational dimensions of customer flow.

3.2.2. Path data

Traffic data offer information only on how many customers enter the store. After entering the store, customers engage with various elements in the store, such as products on shelves and sales personnel. The ways in which customers go about interacting with these store elements provide information about their decision making processes, and hence, the demand. Traditionally, such data have been difficult to gather; however, technologies such as RFID tags, digital video capture, and nimble database solutions have made collecting, storing, coding and accessing such data in a brick-and-mortal retailing context easier. In the context of electronic retailing, such data are available readily, since store websites usually track each visitor's digital clickstream. Such data can be used to assess customer intent, which can then be used for forecasting demand. For example, an electronic retail store can use customer path information to detect customer interest, and to position inventory in (warehouse) locations closer to the customer for fast delivery when such orders materialize.

Over the last decade, a substantial amount of research has utilized path data collected using these novel methods and sources. The main focus of such papers has usually been on obtaining a better understanding of customer behavior; specifically, the ways in which customers interact with various elements on the physical or digital store. For example, Hui, Bradlow, and Fader (2009) and Hui, Fader, and Bradlow (2009) use data collected using RFID tags installed on shopping carts/baskets to verify behavioral hypotheses on customers' purchase processes. They show that customers are more likely to make a purchase when they spend a longer time shopping. In another effort, Lu, Musalem, Olivares, and Schilkrut (2013) measure customers' sensitivity to waiting in a line. Jain, Misra, and Rudi (2016) use video data to quantify the impact of sales assistance on purchase decisions. However, the use of data from such sources for improving demand forecasting has received only a limited amount of attention. Path data are "short-term", and typically are not suitable for planning typical supply chain procurement or replenishment, activities with lead times that can range from a few days to several months. In practice, retailers need to combine path data with data from other sources in order to predict demand (say through econometric models). However, such sophisticated and time-consuming tools may not be accessible to smaller retailers. Thus, future work could focus on the best way of integrating such data into existing easy-to-employ forecasting methods.

3.3. Internet/user-generated content

Last but not least, we consider the impact of data on a customer's behavior prior to a store visit. The key idea here is that such customer activities are indicative of an intent to purchase a given product, and can be used to improve the forecasting process. Traditionally, such data have been difficult to obtain, barring expensive market surveys. However, the explosion of online activities over the last decade, especially the emergence of "social media", and firms' abilities track them, has changed this. Nowadays, firms have access to many tools that allow them not only to see the latest trends, but also to shape them in order to influence customer behavior. Most firms maintain a presence on social media platforms such as Facebook and Twitter, and can use them to interact with their customers directly.

3.3.1. Google trends

Even without an explicit social media presence, though, there are technologies available today that can track customer search behaviors, i.e., patterns of what customers search for on the web, and how often they do it. There is a growing stream of literature on the use of Google Trends, a free tool available from Google, Inc., that tracks customer search data. The trends are reported on a 1–100 scale, where a value of 100 indicates peak popularity of a search term whereas 50 indicates that it is half of its peak value. For example, the popularity of the search term "L. L. Bean" on January 11, 2017, was at a value of 9; but when President Trump tweeted "Buy L. L. Bean" on January 12, the popularity of "L. L. Bean" surged to 100, its peak.

A growing stream of research is finding evidence that the use of Google trends indexes (and those of competing search engines like Baidu) reduces forecast errors, both in- and out-of-sample. The premise is that smaller forecast errors lead to better supply chain performance (see Boone, Ganeshan, Hicks, & Sanders, 2017; Cui, Gallino, Moreno, & Zhang, 2017). Researchers typically use traditional time series forecasts, most often with some form of autoregressive model as a baseline. They then include the search index for specified search terms as a predictor variable when constructing the trend-enhanced forecasting model. The procedure for testing the value of trend variables is to compare the out-of-sample forecast errors of these two models using a rolling window forecast, most commonly with one-step-ahead forecasts. Existing research has shown that Google trends can be used to improve the forecasting of economic indicators such as unemployment rates and benefit claims (Choi & Varian, 2009, 2012; Smith, 2016), GDP (Castle, Fawcett, & Hendry, 2009), and automobile, home and retail sales (Choi & Varian, 2009; Wu & Brynjolfsson, 2015). The Google trends-enhanced economic models typically have out-of-sample MAEs that are anywhere from 4% to 25% smaller than those of the baseline models. Google trend-enhanced models are also common in public health, often for tracking incidences or the spread of disease outbreaks (for example, see Ginsberg et al., 2009, for predicting flu outbreaks and Seifter, Schwarzwalder, Geis, & Aucott, 2010, for tracking Lyme disease).

Trend data have also been shown to be valuable in industries where having real-time data is critical. In financial markets, search query data are used for measuring retail investor attention (Bank, Larch, & Peter, 2011; Da, Engelberg, & Gao, 2011) and market volatility (Dimpfl & Jank, 2016), and for predicting earnings (Da, Engelberg, & Gao, 2014; Drake, Roulstone, & Thornock, 2012). Trend-enhanced models have been used in estimating customer "traffic", for example in tourism for predicting arrivals (Bangwayo-Skeete & Skeete, 2015; Choi & Varian, 2009), hotel bookings (Pan, Chenguang Wu, & Song, 2012; Rivera, 2016), and movie admissions (Hand & Judge, 2012). The MAPE improvements in many of these industry-specific models are between 10% and 40%, raising the possibility that trend data can enhance prediction.

However, research on the application of Google trends data to the prediction of product sales is very limited. The few studies that have tackled this challenge have focused on the forecasting of aggregate-level sales (Choi & Varian, 2009; Fantazzini & Toktamysova, 2015; Nunnari & Nunnari, 2017; Schaer, Kourentzes, & Fildes, 2019), and have shown modest improvements in MAPE (2%–15%) as a result of using Google trends data for enhancing forecasting. We found two studies that investigated whether Google trends data can improve SKU-level forecasts. Through a case study of a food retailer, Boone, Ganeshan, and Hicks (2015) show that Google Trends improve *in-sample* performances by 6%–8%. Later, using five years of data over five SKUs in multiple categories from a specialty food and cookware retailer, (Boone et al., 2017) show *out-of-sample* MAPE improvements of 2%–8%, depending on the SKU category.[8]

While there is considerable enthusiasm regarding the potential for Google trends data to improve forecasting, a lot of work is needed in order to realize this potential. Following the example of Boone et al. (2017), there need to be more studies testing the out-of-sample forecast improvements that can result from the use of search data. Second, there is a need for systematic protocols for generating relevant search terms for a product. There are potentially thousands of terms that may be correlated with target variables — how does one choose which ones to use? In existing studies, researchers have relied on manager intuition (Boone et al., 2017) or used part of a portfolio of terms that are likely to be related to the product. However, these are not scalable for firms that sell thousands of products. Third, Google only indexes terms that meet a certain threshold of traffic, meaning that index values for relevant search terms may simply not be available. In such situations, the researcher has to rely on proxy terms, which adds another source of uncertainty. Fourth, the way in which Google or other search engines calculate the index is a black box. According to Google, the index is based on sampled data, meaning that they are not entirely consistent over time, calling into question the long-term reliability of the use of trend variables. Fifth, search terms may vary with the product lifecycle or with time even for the same target variables; future studies can perhaps address and gain

[8] While only five SKUs were reported, the authors found similar results for over 15 SKUs.

insights from the ways in which changes in search terms impact demand forecasting. Finally, as was also pointed out by Schaer et al. (2019), such models perform poorly in forecasting demand over the longer time frames (even a few days) that are required for supply chain planning. Since Google trends use contemporaneous terms, dynamic forecasting quickly gets very noisy, even over only a few days, which limits these models to very short forecasting horizons. Addressing the ways in which these models can be used strategically with the sales and operations planning process is another a viable area for future research.

3.3.2. Social media

There is a growing stream of literature on ways of harnessing the data created on social media (Facebook, Twitter, YouTube, etc.) in order to provide forecasting insights. The premise is that as more people share their experiences on these platforms, often in real time, it can not only inform the consumers of this data, but also shape how they make decisions. From a technical perspective, the primary focus of most of the academic studies in this arena has been on establishing how such social media entries can be captured and processed into quantitative measures or indices (that capture the "mood" or "sentiment" of these posts). There is a wide range of views as to how these indices should be constructed, including "volume" measures such as the number of posts,[9] the fraction of positive posts, who is posting, the posters' ratings, the number of upvotes, etc.; and "valence" measures, which relate to the context and quality of posts, such as the subject of the post, the "bag of words" that are used in the post, the credibility and influence of the poster, the context in which the post originated, etc. (see Goh et al., 2013, for a review). The aim is to capture the mood related to the topic; or, in a marketing campaign, to determine how product information can be diffused effectively using social media. For example, a potential index for stock market predictions could be the number of positive tweets about Dow or Nasdaq. The creation of these indices is in itself a vibrant area of research. While reviewing such indices is outside the scope of this paper, it provides a rich area of study for the forecasting community. The second step in assessing user-generated content is to augment forecasting models with these social-media measures or indices in order to test whether they improve the forecasts. In addition to linear models, a significant number of research studies that use social media use nonlinear models, typically machine learning methods (with support vector machines and neural networks being those most commonly used).

Twitter data have been used widely for predicting flu outbreaks (Broniatowski, Paul, & Dredze, 2013; Lamb, Paul, & Dredze, 2013; Lampos, De Bie, & Cristianini, 2010), stock prices (Bollen, Mao, & Zeng, 2011; Rao & Srivastava, 2012; Zhang, Fuehres, & Gloor, 2011), box office revenue (Liu, Ding, Chen, Chen, & Guo, 2016; Mishne, Glance, et al., 2006; Sanguinet, 2016), and TV ratings (Wakamiya, Lee, & Sumiya, 2011). Most studies have shown that forecast errors can be improved by adding social media information, but we have not found any studies that investigate how social media platforms impact SKU-level forecasts. Using data from an online retailer, Cui et al. (2017) show how the number and relevance of Facebook entries reduce the out-of-sample MAPEs of aggregate daily sales for an online fashion retailer by 7%–23%. However, they concede that their forecasts hold only for short timeframes. In a controlled study with a premium ice cream brand, Kumar et al. (2013) were able to show that a marketing campaign that is constructed carefully to make use of Facebook can grow sales. However, their study focused on how to spread the "word of mouth" about the product, not specifically on forecasting SKU demand.

While models that incorporate user-generated content show promise for generating sales and revenue, forecasting models specifically have a number of practical limitations in regard to supply chain planning. First, the forecast horizons of models that use such data are often short. In a supply chain context, where lead times tend to be longer than just a few days, short-term search or social media data information cannot be used for demand and inventory planning. However, these models can be useful for the staging of inventory as it is consumed. For example, as a trend or a 'buzz' is detected for a certain SKU in a certain area, contingency plans can be made to move the inventory to that area in order to maximize profits. Second, such models can also be used to make replenishment decisions for consumable goods. For example, as the interest in a particular SKU (as opposed to actual inventory) grows, actions to replenish can be taken earlier than normal in anticipation of that demand. In our opinion, these Internet search and social media variables have their highest impact on short-term 'yield' management scenarios. Connected stores can be fed digital signage based on current sales and trends. These can work in conjunction with beacons and customer devices to identify customers and offer them promotions such as coupons to stimulate sales of selected SKUs or trending items. Flexible digital price tags can change based on the evolution of these contemporaneous Internet variables, giving customers a price break based on their browsing and buying patterns. Customers can also be given the choice of procuring products through other channels, such as other store locations by means of eCommerce, for products that are unavailable in the store they visit. However, making use of contemporaneous Internet and user-generated content depends on how well these variables are integrated into the sales and operations planning and execution processes.

4. Organizational challenges of big data forecasting

4.1. Integrating big data into sales and operations processes

On a strategic level, every firm has to decide whether and how much to incorporate big data technologies into their planning process. This depends on the relative benefits the firm can potentially accrue against the cost of collecting and analyzing such data. Our intent is not to address this strategic question of *whether* and *how much* to integrate big data into organizational processes; rather, we want to highlight the changing supply chain landscape and the challenges of integrating big data specifically for demand and supply chain planning.

[9] We use the word "post" in its very generic form. This could include a tweet or a retweet, or a Facebook entry or a like, or a comment on a public forum, etc.

4.1.1. Capturing big data and connecting it to traditional SOP processes

Big data brings with it the potential to improve product forecasts and give interesting insights into customer behaviors. However, these potential benefits come with enormous practical challenges for demand planners. First, the sheer volume of data can be overwhelming. For example, Walmart collects more than 2.5 petabytes (1 petabyte = 1 million gigabytes) of data every hour from one million customer transactions.[10] However, only about 0.5% of all data collected is analyzed.[11] Thus, a practical question is, what data should be stored, and for how long? Second, Feng and Shanthikumar (2018) point out that, while "theoretically more information leads to better forecasts, the challenge, however, comes from dealing with the increased number of variables and their ambiguous relationships." Since big datasets, especially those used in forecasting, tend to be sparse and non-repetitive, they posit that semi- or non-parametric methods such as machine learning (see for example Cui et al., 2017) are better suited for analyzing them than traditional time series forecasting methods. Demand planners need to adjust so as to accommodate a wider range of methodologies into the planning process.

4.1.2. From human judgement to data-driven decisions

Studies of industry practices repeatedly show that judgmental forecasts are the norm and are based on "gut feelings." Here, statistical forecasts are adjusted to account for numerous factors that demand planners perceive as being difficult to measure, including promotional activities, seasonal activities, demand risk, demand and supply chain disruptions (see for example Fildes & Goodwin, 2007). While human judgement has the potential to improve forecasts, experts often introduce their own biases in the process (see for example Hypotheses H_4 and H_5 of Fildes, Goodwin, Lawrence, & Nikolopoulos, 2009). Often, such judgmental forecasts have an adverse impact on the forecast accuracy, since they weight their contribution disproportionately (Franses & Legerstee, 2010). At least in theory, newly available data have the potential to reduce the negative effects of "adjustments" to the forecast (see the prototypical models of Feng & Shanthikumar, 2018), and there have been efforts to integrate such new data streams (Sagaert, Aghezzaf, Kourentzes, & Desmet, 2018) into forecasts in a systemic manner. However, significant practical challenges remain. First, recent surveys (Weller & Crone, 2012) indicate that such new data streams may not be available readily, or be used by a majority of demand planners. Second, significant hardware, software, and analytical support is required in order to integrate the data into the ERP systems that planners typically use. Third, a significant learning curve is required in order to interpret the results from the newer machine learning algorithms. Another area of research for the forecasting community is to investigate whether big data streams can substitute for expert judgement. While we do not believe that big data can ever supplant expert opinion, it can alleviate some of the issues with judgmental forecasts.

[10] see https://www.dezyre.com/article/how-big-data-analysis-helped-increase-walmarts-sales-turnover/109.

[11] https://www.technologyreview.com/s/514346/the-data-made-me-do-it/.

4.1.3. Changing customer experiences

The widespread use of connected devices such as smartphones or wearable technologies are changing the in-store customer experience. Beacons in stores (or online) can detect a customer entering a store, thus providing significant opportunities for customer engagement, gathering data and making links to historical buying patterns. As a result, demand is not just forecast, but can be influenced and modified directly, as consumer buying can now be directed (see ; Cohen, 2018; Feng & Shanthikumar, 2018; and the various references within). Content such as product information, pricing, and promotions such as e-coupons can be personalized to the customer (and sent to their phones) based on their past browsing experience (Ganeshan, 2014). Products or substitutes can be suggested, and in some cases, customers can be led to product locations. This is an area of active research in the marketing and operations management communities (see for example Cachon, Daniels, & Lobel, 2017; Chen, Ma, Simchi-Levi, & Xin, 2016, for pricing; van Heerde & Neslin, 2017, for promotions; Bobadilla, Ortega, Hernando, & Gutiérrez, 2013, for recommendation systems; and Lau, Zhang, & Xu, 2017, for sentiment analysis).

From a customer's perspective, smart devices are helpful for researching a product (for example, looking for reviews), looking up a competitor's price or even ordering a competing product right from the store. In many ways, the customer experience has become a very personalized one. Demand planners, and indeed the research community, need to address this apparent contradiction of disaggregate planning for the unique customer at significantly shorter time horizons while at the same time planning for aggregate needs over the longer-term. Second, customers are now demanding the flexibility of being able to order the product via multiple channels (the "omni-channel" experience), and want product delivery (and potential returns) in a wide variety of locations, including their home, stores, or third-party locations. While big data technologies have enabled and enhanced the planning between channels, it still requires a significant amount of coordination between channels (eCommerce, retail, etc.), in terms not only of tactical planning, but also of shared strategic goals, performance measures, and incentives. Indeed, forecasting omni-channel demand remains a viable area of enquiry.

4.1.4. Integrating the connected supply chain

Big data technologies have enabled the echelons of the supply chain to be "connected" and to "communicate" with one another, creating the "intelligent" supply chain Sanders (2016). At the customer level, there has been a profusion of "connected" devices. This has enabled a more active engagement with the customer. Targeted messages/promotions can reach the customer's living room; consumption can be tracked to actuate replenishment; problems can be diagnosed online; and finally, specialized products and services can be tailored to the customer. Such connectivity provides a treasure trove of data that the demand planner can potentially use to capture the customers' intent, to shape demand, and to increase loyalty. It can also provide a greater supply chain responsiveness to forecast errors. However, in our experience, demand

planners seldom use this data for the S & OP process. The integration of customer engagement provides a viable field of enquiry for the forecasting community.

Initiatives on the supply side, such as collaborative planning, forecasting and replenishment (CPFR) and efficient customer response (ECR), are setting up processes where demand planning and replenishment is done in collaboration with key suppliers. Events such as store openings, promotions, and supply shortages are shared between trading partners, improving visibility and forecast accuracy. In many cases, firms also extend this visibility to the customer (Boone & Ganeshan, 2008; Panahifar, Heavey, Byrne, & Fazlollahtabar, 2015), especially in regard to inventory availability. RFID tags help to improve inventory tracking and maintain consistent service levels (Bertolini, Bottani, Rizzi, & Volpi, 2010). Emerging technologies such as Blockchains promise secure, accurate transactions and easy reporting. They also make "track and trace" in the supply chain easier, providing the demand planner with valuable information on planning replenishment and order-promise dates. Finally, instrumented supply chains have improved yield management (for example by using dynamic price tags) and supply chain efficiency. The challenge for demand planners is how these new technologies should be integrated into the planning process. For example, should Blockchain technology be rolled out to all products? Should they embrace smart contracts? How should they establish the chain of command? Answering these questions successfully will enable an improved visibility and potentially better forecasts.

4.2. Privacy, bias, and discrimination

Connected devices and sensors are constantly collecting significant amounts of data on customers. Often, such data (such as mobile phone pings, for example) do not have any specific purpose. In addition, the data are not anonymous: often the identity of the consumer is linked to the behavior. Since the forecasting of consumer behavior often warrants the use of these private datasets, it raises significant issues of privacy and data integrity, as well as bias and discrimination, as the ensuing discussion will show. These issues are increasing in prominence,[12] and the forecasting community needs to be actively engaged in finding solutions.

The Identity Theft Resource Center estimates[13] that there have been 8190 data breaches since 2005, with over one billion customer records having been exposed to theft. These breaches can have disastrous consequences, both for the firms whose data were hacked and potentially for the consumers whose identities were exposed.[14] In 2017, the largest breach was Equifax, a credit monitoring agency in the US. Over 140 million US and half a million British customer records — names, addresses, and social security numbers — were hacked in this breach. At best, the cost to the consumer is the cost of protecting their identity; at worst, it is the cost of reversing the consequences of their identity being stolen.

Big data are also not immune to the typical problems faced by traditional datasets, such as that of bias and representation in datasets. The *device* or *sensor* collecting the data can introduce bias into the data. For example, the City of Boston used the "Street Bump" App[15] to collect data from citizens on potholes. The user's smartphone's accelerometer would record "bumps" and their locations as they drove along on the road, a potential sign of a pothole that the city could then fix. The app reported disproportionate numbers of potholes in wealthier parts of town where more residents owned smartphones and were digitally engaged. Thus, even such well-intentioned programs have the ability to have a negative impact on members of our society.

Another famous example is the overestimation of flu incidence in algorithms that used Google Trends search terms, due to people beginning searching for flu symptoms when they heard of flu incidences on the news. The CDC data were field-based and tended to estimate flu incidences better. These are examples of so-called "big data hubris", the implicit assumption that big data collected via automated or volunteered methods can replace traditional data collection and analysis (Lazer, Kennedy, King, & Vespignani, 2014).

4.2.1. Algorithmic ethics and injustice

For a recent job advertisement in finance located in Washington DC, Verizon, a US telecom company, targeted their promotion at "the Facebook feeds of users 25 to 36 years old who lived in the nation's capital, or had recently visited there, and had demonstrated an interest in finance."[16] Such targeted ads raise questions of fairness to older workers, and many critics suggested that the ad violated the federal Age Discrimination in Employment Act of 1967. Such biases by algorithms based on demographic features (race[17] and disability[18] are also cited commonly) have come to be known as "machine bias". O'Neil's (2017) best-selling book *Weapons of Math Destruction* provides multiple examples of such algorithmic injustice — poor evaluation procedures for high school teachers, race, gender and economic biases in product offerings, etc. — and these algorithmic decisions often affect the most vulnerable populations.

---

12 Not only are companies, especially retailers, reacting to many breaches and the algorithmic bias discussed in this section, policy makers are also taking steps to legislate against privacy- and algorithmic-injustice. See: https://www.cnet.com/news/elizabeth-warren-equifax-mark-warner-credit-reporting-agencies-data-breach-bill-fines/.

13 https://www.idtheftcenter.org/Data-Breaches/data-breaches.

14 See for example the story of Drew Armstrong, whose identity was stolen: https://www.bloomberg.com/news/articles/2017-09-13/my-three-years-in-identity-theft-hell.

15 John Podesta, President Barack Obama's senior counselor, references this bias in an interview. See https://www.reuters.com/article/us-usa-obama-privacy/white-house-looks-at-how-big-data-can-discriminate-idUSBREA3Q00M20140427.

16 https://www.propublica.org/article/facebook-ads-age-discrimination-targeting.

17 Propublica reported that Facebook was allowing housing advertisers to exclude viewers by race. See: https://www.propublica.org/article/facebook-advertising-discrimination-housing-race-sex-national-origin. Amazon also came under criticism when it rolled out Amazon same day deliveries in cities, as their algorithms inadvertently excluded black neighborhoods.

18 O'Neil (2017) describes how job applicants were weeded out based on mental health.

Several research streams tackle the issue of machine bias (see for example Calders & Verwer, 2010). These are typically statistical methods that indicate potential biases among customer attributes.

As the use of big data streams grows increasingly widespread in the forecasting research community, we expect this to become a important research area. The forecasting research community, with its expertise in statistical methods, can contribute significantly to (a) fraud detection (i.e., detecting outliers in data streams) in big data; (b) determining how "clean"[19] internet and user-generated content can be used in forecasting models; and (c) establishing how big data or machine algorithms can be bias-free (for example, generating "bias" scores for forecasting models that predict customer behavior).

5. Summary and conclusions

Our intent in this paper has been to provide an overview of the way in which the forecasting of sales in the supply chain could be enhanced by customer analytics based on big data and associated technologies. We have focused rather broadly on three aspects of the purchase decision: the research phase, the store experience, and finally the sales transaction. From this perspective, we have reviewed the many ways in which big data can enhance *aggregate* forecasts. First, we show how the granular data that are now available through POS systems can help to overcome demand censuring and provide a holistic view encompassing a multi-product perspective. Second, in-store technologies can be used to assess traffic and customer purchasing behaviors, which can inform aggregate sales forecasts. Finally, we show how researchers have used user-generated data from Internet searches and social media to enhance forecasting models.

While the use of big data shows promise, it also introduces significant challenges. First, the size and unstructured content of these enormous data sets can be daunting. Second, the data are sparse and non-repetitive and requires multiple methodologies for forecasting, which slows their widespread acceptance. Finally, customer experiences and supply chains are changing, making big data technologies difficult to integrate into conventional SOP processes. We concur with the assessment of the Associate Editor who reviewed this paper: "As yet, in demand forecasting we would see 'big data' mostly as potential with gains yet to be delivered".

Firms that embrace big data technologies must also address three key related management issues: privacy, security, and governance. First, since a multitude of sensors are used to gather information about customer browsing and purchase behaviors, it is imperative that firms address and articulate a clear privacy policy. The growth of technologies is outpacing privacy laws in many jurisdictions (Machanavajjhala & Reiter, 2012), which adds the challenge that firms potentially may have to deal with privacy laws in different geographical locations. Second, these large datasets (especially if they are distributed) pose a security risk (Lu, Zhu, Liu, Liu, & Shao, 2014). Any failure to keep customer data, especially sensitive data (such as social security and credit card numbers) secure can be illegal and open to relevant government oversight agencies. Finally, firms that make use of big data also need a clear big data strategy in order to keep their decisions free of bias. For example, a multidisciplinary group of researchers known as the Fairness, Accountability and Transparency in Machine Learning (FAT/ML) community lay out a set of five principles that can potentially alleviate issues of bias: responsibility, explainability, accuracy, auditability, and fairness. The aim of these principles is to "help developers and product managers design and implement algorithmic systems in publicly accountable ways. Accountability in this context includes an obligation to report, explain, or justify algorithmic decision-making as well as mitigate any negative social impacts or potential harms".[20]

A clearly articulated governance structure will lay out what data are collected where and from whom; the way in which these data and the accompanying decisions' quality are validated; and how this data can be leveraged as an organizational asset (see for example Otto, 2011, Hashem et al., 2015).

Despite the challenges, we see the next decade as being the golden age of forecasting. We see a future where demand planners can make use of large and varied datasets in real-time; can provide a clearer picture of customer behavior; and can generate accurate forecasts that will help not only to improve the efficiency of their supply chains but also to enhance their revenues. We hope that our optimism is shared by the forecasting community — we are looking forward to the increased interest in big data technologies and the development of new methods and tools for turning such data into more accurate forecasts for use in making actionable operational decisions.

References

Bangwayo-Skeete, P. F., & Skeete, R. W. (2015). Can Google data improve the forecasting performance of tourist arrivals? Mixed-data sampling approach. *Tourism Management*, *46*, 454–464.

Bank, M., Larch, M., & Peter, G. (2011). Google search volume and its influence on liquidity and returns of German stocks. *Financial Markets and Portfolio Management*, *25*(3), 239–264.

Bertolini, M., Bottani, E., Rizzi, A., & Volpi, A. (2010). The benefits of RFID and EPC in the supply chain: Lessons from an Italian pilot study. In *The internet of things* (pp. 293–302). Springer.

Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-based Systems*, *46*, 109–132.

Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computer Science*, *2*(1), 1–8.

Boone, T., & Ganeshan, R. (2008). The value of information sharing in the retail supply chain: two case studies. *Foresight: The International Journal of Applied Forecasting*, *9*, 12–17.

Boone, T., Ganeshan, R., & Hicks, R. L. (2015). Incorporating Google trends data into sales forecasting. *Foresight: The International Journal of Applied Forecasting*, (38), 9–14.

Boone, T., Ganeshan, R., Hicks, R. L., & Sanders, N. R. (2017). Can Google trends improve your sales forecast?. *Production and Operations Management*, *27*(10), 1770–1774.

Broniatowski, D. A., Paul, M. J., & Dredze, M. (2013). National and local influenza surveillance through Twitter: an analysis of the 2012–2013 influenza epidemic. *PLoS One*, *8*(12), e83672.

[19] By this we mean that the content being captured correlates with the signal being measured.

[20] https://www.fatml.org/resources/principles-for-accountable-algorithms, accessed January 15, 2018.

Bruno, H., & Vilcassim, N. (2008). Structural demand esimation with varying product availability. *Marketing Science*, *27*(6), 1126–1131.

Cachon, G. P., Daniels, K. M., & Lobel, R. (2017). The role of surge pricing on a service platform with self-scheduling capacity. *Manufacturing and Service Operations Management*, *19*(3), 368–384.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Castle, J. L., Fawcett, N. W., & Hendry, D. F. (2009). Nowcasting is not just contemporaneous forecasting. *National Institute Economic Review*, *210*(1), 71–89.

Chen, X., Ma, W., Simchi-Levi, D., & Xin, L. (2016). *Dynamic recommendation at checkout under inventory constraint, MIT working paper*.

Chen, L., Mersereau, A. J., & Wang, Z. (2017). Optimal merchandise testing with limited inventory. *Operations Research*, *65*(4), 968–991.

Choi, H., & Varian, H. (2009). *Predicting initial claims for unemployment benefits*. Google Inc.

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*(s1), 2–9.

Chung, J., & Rao, V. R. (2003). A general choice model for bundles with multiple-category products: application to market segmentation and optimal pricing for bundles. *Journal of Marketing Research*, *40*(2), 115–130.

Cohen, M. C. (2018). Big data and service operations. *Production and Operations Management*, *27*(9), 1709–1723.

Cui, R., Gallino, S., Moreno, A., & Zhang, D. J. (2017). The operational value of social media information. *Production and Operations Management*, *27*(10), 1749–1769.

Da, Z., Engelberg, J., & Gao, P. (2011). In search of attention. *The Journal of Finance*, *66*(5), 1461–1499.

Da, Z., Engelberg, J., & Gao, P. (2014). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies*, *28*(1), 1–32.

Dimpfl, T., & Jank, S. (2016). Can internet search queries help to predict stock market volatility? *European Financial Management*, *22*(2), 171–192.

Drake, M. S., Roulstone, D. T., & Thornock, J. R. (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research*, *50*(4), 1001–1040.

Fantazzini, D., & Toktamysova, Z. (2015). Forecasting German car sales using Google data and multivariate models. *International Journal of Production Economics*, *170*, 97–135.

Feng, Q., & Shanthikumar, J. G. (2018). How research in production and operations management may evolve in the era of big data. *Production and Operations Management*, *27*(9), 1670–1684.

Fildes, R., & Goodwin, P. (2007). Against your better judgment? How organizations can improve their use of management judgment in forecasting. *Interfaces*, *37*(6), 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*(1), 3–23.

Franses, P. H., & Legerstee, R. (2010). Do experts' adjustments on model-based SKU-level forecasts improve forecast quality?. *Journal of Forecasting*, *29*(3), 331–340.

Ganeshan, R. (2014). Clickstream analysis for forecasting online behavior. *Foresight: The International Journal of Applied Forecasting*, (33), 15–19.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012–1014.

Goh, K. -Y., Heng, C. -S., & Lin, Z. (2013). Social media brand community and consumer behavior: Quantifying the relative impact of user- and marketer-generated content. *Information Systems Research*, *24*(1), 88–107.

Hand, C., & Judge, G. (2012). Searching for the picture: forecasting UK cinema admissions using Google Trends data. *Applied Economics Letters*, *19*(11), 1051–1055.

Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, S. U. (2015). The rise of "big data" on cloud computing: Review and open research issues. *Information Systems*, *47*, 98–115.

Hruschka, H. (2017). Multicategory purchase incidence models for partitions of product categories. *Journal of Forecasting*, *36*(3), 230–240.

Hui, S., Bradlow, E., & Fader, P. (2009). Testing behavioral hypotheses using an integrated model of grocery store shopping path and purchase behavior. *Journal of Consumer Research*, *36*(3), 478–493.

Hui, S., Fader, P., & Bradlow, E. (2009). Research note — The traveling salesman goes shopping: The systematic deviations of grocery paths from TSP optimality. *Marketing Science*, *28*(3), 566–572.

Jain, A., Misra, S., & Rudi, N. (2016). *Sales Assistance, Search and Purchase Decisions: an Analysis Using Retail Video Data, Working Paper*. CUNY, NY: Ziklin School of Business.

Jain, A., Rudi, N., & Wang, T. (2015). Demand estimation and ordering under censoring: stock-out timing is (almost) all you need. *Operations Research*, *63*(1), 134–150.

Karabati, S., Tan, B., & Öztürk, Ö. C. (2009). A method for estimating stock-out-based substitution rates by using point-of-sale data. *IIE Transactions*, *41*(5), 408–420.

Kumar, V., Bhaskaran, V., Mirchandani, R., & Shah, M. (2013). Creating a measurable social media marketing strategy: Increasing the value and ROI of intangibles and tangibles for hokey pokey. *Marketing Science*, *32*(2), 194–212.

Lam, S. Y., Vandenbosch, M., Hulland, J., & Pearce, M. (2001). Evaluating promotions in shopping environments: Decomposing sales response into attraction, conversion, and spending effects. *Marketing Science*, *20*(2), 194–215.

Lam, S., Vandenbosch, M., & Pearce, M. (1998). Retail sales force scheduling based on store traffic forecasting. *Journal of Retailing*, *74*(1), 61–88.

Lamb, A., Paul, M. J., & Dredze, M. (2013). Separating fact from fear: Tracking flu infections on Twitter. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: Human language technologies* (pp. 789–795).

Lampos, V., De Bie, T., & Cristianini, N. (2010). Flu detector—tracking epidemics on Twitter. In *Joint European conference on machine learning and knowledge discovery in databases* (pp. 599–602). Berlin, Heidelberg: Springer.

Lau, R. Y. K., Zhang, W., & Xu, W. (2017). Parallel aspect-oriented sentiment analysis for sales forecasting with big data. *Production and Operations Management*, *27*(10), 1775–1794.

Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: traps in big data analysis. *Science*, *343*(6176), 1203–1205.

Liu, T., Ding, X., Chen, Y., Chen, H., & Guo, M. (2016). Predicting movie box-office revenues by exploiting large-scale social media content. *Multimedia Tools and Applications*, *75*(3), 1509–1528.

Lu, Y., Musalem, A., Olivares, M., & Schilkrut, A. (2013). Measuring the effect of queues on customer purchases. *Management Science*, *59*(8), 1743–1763.

Lu, R., Zhu, H., Liu, X., Liu, J. K., & Shao, J. (2014). Toward efficient and privacy-preserving computing in big data era. *IEEE Network*, *28*(4), 46–50.

Machanavajjhala, A., & Reiter, J. P. (2012). Big privacy: protecting confidentiality in big data. *XRDS: Crossroads, The ACM Magazine for Students*, *19*(1), 20–23.

Mani, V., Kesavan, S., & Swaminathan, J. M. (2015). Estimating the impact of understaffing on sales and profitability in retail stores. *Production and Operations Management*, *24*(2), 201–218.

Mishne, G., Glance, N. S., et al. (2006). Predicting movie sales from blogger sentiment. In *AAAI Spring symposium: Computational approaches to analyzing weblogs* (pp. 155–158).

Musalem, A., Olivares, M., Bradlow, E., Terwiesch, C., & Corsten, D. (2010). Structural estimation of the effect of out-of-stocks. *Management Science*, *56*(7), 1180–1197.

Nunnari, G., & Nunnari, V. (2017). Forecasting monthly sales retail time series: A case study. In *2017 IEEE 19th Conference on Business Informatics, vol. 1* (pp. 1–6). IEEE.

O'Neil, C. (2017). *Weapons of math destruction: how big data increases inequality and threatens democracy*. Broadway Books.

Otto, B. (2011). Organizing data governance: Findings from the telecommunications industry and consequences for large service providers. *Communications of the Association for Information Systems*, *29*, 3.

Pan, B., Chenguang Wu, D., & Song, H. (2012). Forecasting hotel room demand using search engine data. *Journal of Hospitality and Tourism Technology*, *3*(3), 196–210.

Panahifar, F., Heavey, C., Byrne, P. J., & Fazlollahtabar, H. (2015). A framework for collaborative planning, forecasting and replenishment (CPFR) state of the art. *Journal of Enterprise Information Management*, *28*(6), 838–871.

Perdikaki, O., Kesavan, S., & Swaminathan, J. (2012). Effect of traffic on sales and conversion rates of retail stores. *Manufacturing and Service Operations Management*, *14*(1), 145–162.

Queenan, C., Ferguson, M., Higbie, J., & Kapoor, R. (2007). A comparison of unconstraining methods to improve revenue management systems. *Production and Operations Management*, *16*(6), 729–746.
Rao, T., & Srivastava, S. (2012). Analyzing stock market movements using Twitter sentiment analysis. In *Proceedings of the 2012 international conference on advances in social networks analysis and mining* (pp. 119–123). IEEE Computer Society.
Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management*, *57*, 12–20.
Russell, G. J., & Petersen, A. (2000). Analysis of cross category dependence in market basket selection. *Journal of Retailing*, *76*(3), 367–392.
Sagaert, Y. R., Aghezzaf, E. -H., Kourentzes, N., & Desmet, B. (2018). Temporal big data for tactical sales forecasting in the tire industry. *Interfaces*, *48*(2), 121–129.
Sanders, N. R. (2016). How to use big data to drive your supply chain. *California Management Review*, *58*(3), 26–48.
Sanguinet, M. E. (2016). *Hashtags, tweets and movie receipts: Social media analytics in predicting box office hits* (Ph.D. thesis), San Diego State University.
Schaer, O., Kourentzes, N., & Fildes, R. (2019). Demand forecasting with user-generated online information. *International Journal of Forecasting*, *35*(1), 197–212.
Seifter, A., Schwarzwalder, A., Geis, K., & Aucott, J. (2010). The utility of "Google Trends" for epidemiological research: Lyme disease as an example. *Geospatial health*, *4*(2), 135–137.
Smith, P. (2016). Google's MIDAS touch: Predicting UK unemployment with internet search data. *Journal of Forecasting*, *35*(3), 263–284.
Tan, B., & Karabati, S. (2004). Can the desired service level be achieved when the demand and lost sales are unobserved?. *IIE Transactions*, *36*(4), 345–358.
van Heerde, H. J., & Neslin, S. A. (2017). Sales promotion models. In *Handbook of marketing decision models* (pp. 13–77). Springer.
van Ryzin, G., & Vulcano, G. (2014). A market discovery algorithm to estimate a general class of nonparametric choice models. *Management Science*, *61*(2), 281–300.
Wakamiya, S., Lee, R., & Sumiya, K. (2011). Towards better TV viewing rates: exploiting crowd's media life logs over Twitter for TV rating. In *Proceedings of the 5th international conference on ubiquitous information management and communication* (p. 39). ACM.
Wecker, W. (1978). Predicting demand from sales data in the presence of stockouts. *Management Science*, *24*(10), 1043–1054.
Weller, M., & Crone, S. (2012). *Supply chain forecasting: Best practices and benchmarking study*: *Technical paper*, Lancaster Centre for Forecasting.
Wu, L., & Brynjolfsson, E. (2015). The future of prediction: How Google searches foreshadow housing prices and sales. In *Economic Analysis of the Digital Economy* (pp. 89–118). University of Chicago Press.
Zhang, X., Fuehres, H., & Gloor, P. A. (2011). Predicting stock market indicators through Twitter "I hope it is not as bad as I fear". *Procedia-Social and Behavioral Sciences*, *26*, 55–62.

Tonya Boone is an Associate Professor at the Raymond A. Mason School of Business, The College of William & Mary. Prior to joining the faculty of the College of William and Mary, she was a faculty member at the Fischer College of Business at Ohio State University.

Her research and teaching interests include sustainable operations, knowledge management in professional service organizations, and the management of supply chains in data rich environments. Tonya's research has been published in the top academic management journals, including *Management Science, Journal of Operations Management, POMS Journal*, and *Decision Sciences*. She is also a co-editor of the recent book *Sustainable Supply Chain Management: Methods, Models, and Policy Implications*.

Tonya is active in her community, serving on the Board of the Williamsburg Economic Development Authority, the Board of the Hampton Roads Incubator, and the Board of the Williamsburg Regional Library. She is a sought-after expert in the local community in the areas of sustainability, public service, and supply chain management.

Tonya Boone has a Ph.D. in Operations and Technology Management from the University of North Carolina at Chapel Hill's Kenan-Flagler School of Business; a MBA from the College of William and Mary; and a B.S. in Electrical & Electronics Engineering from the University of Kansas.

Ram Ganeshan is the D. Hillsdon Ryan Professor of Business at the Raymond A. Mason School of Business, The College of William and Mary, Williamsburg, VA.

Ram's teaching, research and consulting interests are in the areas of supply chain management, data analytics, and logistics strategy, primarily in the chemical, hi-tech, and retail industries. He is a regular contributor to academic and trade journals and is the co-editor of three books including *Quantitative Models for Supply Chain Management* which is one of the most highly cited books in supply chain management. In 2001, the Production & Operations Management Society (POMS) awarded him the prestigious Wickham Skinner Award for his research on how supply chains can be efficiently managed. His current research projects extensively rely on big-data techniques to provide insights into making sound decisions for managing both manufacturing and service supply chains. They include: (1) work with online retailers to harness and interpret clickstream data from customer browsing behavior; (2) working with carriers in the transportation sector to reduce cost and carbon footprint; and (3) analyzing large scale project data in professional services to provide strategies for productivity improvement.

He received a Doctorate in Operations and Logistics Management from Penn State; a MSOR degree in Operations Research from the University of North Carolina at Chapel Hill; and an undergraduate degree in Industrial Management from the Birla Institute of Technology & Science in India.

Aditya Jain is assistant professor of Operations Management at Zicklin Business School at Baruch College in New York city. His primary research interests include retail operations, supply chain management and healthcare operations. Aditya earned his Ph.D. in Operations Management from the Simon School of Business at University of Rochester, and served as a faculty member at Indian School of Business and the Kellogg School of Northwestern University prior to his current appointment. His industry experience includes extensive consulting work with fashion retailer through the retail/supply chain analytics company he founded and ran for several years. He holds bachelor's degree in Chemical Engineering from Indian Institute of Technology, Bombay.

Nada R. Sanders is the Distinguished Professor of Supply Chain Management at the D'Amore-McKim School of Business at Northeastern University in Boston MA. Prior to that, she held the Iacocca Chair at the College of Business and Economics at Lehigh University and as the West Chair at the M.J. Neeley School of Business.

Her research and teaching interests have been in forecasting and the use of data analytics in decision-making within the supply chain context. She has authored over one hundred scholarly works and has served on the editorial boards of prominent journals in her field, including the *Journal of Operations Management, Production and Operations Management, Decision Sciences Journal, Journal of Business Logistics*, and *International Institute of Forecasting*. She is a Fellow of the *Decision Sciences Institute* and was co-founder and Associate Editor of *Foresight: The International Journal of Applied Forecasting*. She has authored multiple books with the most recent being Big Data Driven Supply Chain Management and has given numerous talks on the subject including a recent HBR webinar. She currently serves on the Board of POMS, having served as both Program and General Chair.

She holds a Ph.D. in Operations Management and Logistics, and an MBA, from the Fisher College of Business at The Ohio State University, as well as a B.S. in Mechanical Engineering.