POMS PRODUCTION AND OPERATIONS MANAGEMENT

Production and Operations Management 25(12), pp. 2002–2013, © 2016 Production and Operations Management Society

Inventory Management in the Era of Big Data

Dimitris Bertsimas
MIT Sloan School of Management and Operations Research Center, 77 Massachusetts Ave., E40-111, Cambridge, MA 02139, USA, dbertsim@mit.edu

Nathan Kallus
Cornell University and Cornell Tech, 111 Eighth Ave #302 New York, NY 10011, USA, kallus@cornell.edu

Amjad Hussain
Silkroute, 950 Stephenson Hwy Troy, Michigan 48083 USA, amjad.hussain@silkrouteglobal.com

1. Introduction

The explosion in the availability and accessibility of machine-readable data is creating new opportunities for better decision making in applications of operations management. The swell of data and advances in machine learning have enabled applications that predict, for example, consumer demand for video games based on online web-search queries (Choi and Varian 2012) or box-office ticket demand based on Twitter chatter (Asur and Huberman 2010). In the context of inventory management, demand is the key uncertainty affecting decisions and such works suggest a potential opportunity to leverage large-scale web data to improve inventory decisions, for example, for stocking video game titles or allocating cinemas of varying capacities. There are also many other applications of machine learning, including Da et al. (2011), Goel et al. (2010), Gruhl et al. (2004, 2005), Kallus (2014), that use large-scale and web-based data to generate predictions of quantities that may in fact be of interest in operations management applications. By and large, however, these applications and the machine learning techniques employed do *not* address optimal decision-making under uncertainty that is appropriate for operations management problems and, in particular, for inventory management.

We study how these data, leveraged appropriately, can correctly and successfully inform inventory management decisions and provide a competitive edge. We focus on a particular case study of the distribution and manufacturing arm of a global media conglomerate (henceforth, the vendor), which, as a distributor of multi-media, is among the three largest in the world. The vendor, which shall remain unnamed, is a direct customer of Silkroute, a provider of analytics platforms for managing manufacturing, distribution, and retail operations. The vendor, which ships an average of 1 billion units in a year, as well as the media retail industry at large, is under increased pressure to improve operations and lower costs in the face of increasing digitalization, declining sales, and diminishing shelf space. The heightened importance and consequence of good inventory decisions provide an excellent case study of the use of large-scale data for achieving a competitive edge in a squeezed industry.

We consider the vendor's VMI (vendor-manage inventory) operations in selling over half-a-million entertainment titles on CD, DVD, and BluRay at major European retailers with over 20,000 locations. To inform VMI decisions, we leverage transactional records collected and organized by the Silkroute platform, data we harvested from public Internet sources including IMDb.com (International Movie Database) and RottenTomatoes.com, and search query volume data provided by Google Trends.

To leverage these data, we employ recent data-driven optimization techniques developed by Bertsimas and Kallus (2014) that address the *conditional stochastic optimization problem*:

$$\begin{aligned} v^*(x) &= min_{z\epsilon\mathbb{Z}}\mathbb{E}[c(z;Y)|X=x], \\ z^*(x) &\in argmin_{z\epsilon\mathbb{Z}}\mathbb{E}[c(z;Y)|X=x], \end{aligned} \tag{1}$$

wherein, on the basis of an observation of auxiliary covariates $X \in \mathbb{R}^d x$, a decision $z(x)$, constrained in a feasible space $\mathbb{Z}\subset\mathbb{R}^d z$, is chosen in an optimal manner to minimize an uncertain cost $c(z; Y)$ that depends on a random variable $Y \in \mathbb{R}^d y$. For example, in the context of media retail inventory management, the uncertain quantities Y of direct impact on costs are the demands for stocked products; the decisions are quantities $z \geq 0$ for each product, constrained by limited capacity $1^T z \leq K$; and, the auxiliary covariates X that may help us choose the best quantities may include recent sale figures, volume of Google searches for a products or company, news coverage, or user reviews. The solution $z^*(x)$ to problem (1) represents the full-information optimal decision, which, via full knowledge of the joint dis-

tribution of X, Y, leverages the observation $X = x$ to the fullest possible extent in minimizing expected costs. In practice, the underlying joint distribution of X, Y is unknown and we must devise a policy $\hat{z}_N(x)$ based only on data $S_N = \{(x^1, y^1), \ldots, (x^N, y^N)\}$. This learning task was addressed in Bertsimas and Kallus (2014), where new methods for this problem are developed, which have two important properties:

Asymptotic optimality:

$$lim_{N\to\infty}\mathbb{E}[c(\hat{z}_N(x); Y)|X = x] = v^*(x)$$

for almost everywhere x, almost surely.

Tractability: $\hat{z}_N(x)$ can be computed in polynomial time and oracle calls, and, in many important cases, it is solvable using off-the-shelf optimization solvers.

One of the simplest approaches proposed in Bertsimas and Kallus (2014) is based on k-nearest neighbors (kNN), where we let

$$\hat{z}_N^{kNN}(x) \in \text{argmin}_{z\in\mathcal{Z}} \sum_{i\in N_k(x)} c(z; y^i),$$

$$N_k(x) = \{i : x^i \text{ is among the } kNN\text{s to x in the data}\}.$$

Various additional methods are developed in Bertsimas and Kallus (2014). The coefficient of prescriptiveness is defined in Bertsimas and Kallus (2014) as

$$P = \frac{\mathbb{E}[c(\hat{z}_N(x); Y)] - \mathbb{E}[\min_{z\in\mathcal{Z}} c(z; Y)]}{\min_{z\in\mathcal{Z}}\mathbb{E}[c(z; Y)] - \mathbb{E}[\min_{z\in\mathcal{Z}} c(z; Y)]},$$

which unitlessly measures the prescriptive content of the auxiliary data X and the efficacy of the policy $\hat{z}_N(x)$ with respect to operational costs in a manner analogous to coefficient of determination R^2 for prediction. In our case study, the rich, large-scale data collected combined with these advances in data-driven optimization account for an 88% reduction in operational costs as measured by P. That is, our approach, based on the data we collect and the prescriptive algorithms we use, takes reduces 88% of excess costs due to uncertainty – a significant advance in addressing the industry's emerging challenges.

2. Problem Description and Formulation

The retail locations in the VMI network range from electronic home goods stores to supermarkets, gas stations, and convenience stores. Under VMI, what is sold at the locations and its replenishment (which is performed weekly) is managed by the vendor. Procurement is done under scan-based trading (SBT), which means that the vendor owns all inventory until scanned at point-of-sale, at which point the retailer procures the unit and sells to the customer. This means that retailers have no cost of capital in holding the vendor's inventory. The cost of a unit is driven primarily by the fixed cost of content production; manufacturing (pressing) media and delivery costs are secondary. Therefore, maximizing network-wide sell-through is the primary objective of the vendor. The limiting factor is capacity: there is limited shelf space (often limited to an aisle endcap display) and generally no storage. Thus, the main loss incurred in over-stocking a particular product lies in the loss of potential sales of another product that sold out (or was not stocked at all) but could have sold more, and there are many potential products. Apart from the limited shelf space, the other primary difficulty is the high uncertainty inherent in the initial demand for new releases, which, at the same time, drive the most sales.

Let $r = 1, \ldots, R$ index the locations, $t = 1, \ldots, T$ index the replenishment periods, and $j = 1, \ldots, d$ index the products. Denote by z_j the order quantity decision for product j, by Y_j the uncertain demand for product j, and by K_r the overall inventory and display capacity at location r. Optimizing sell-through as discussed in the previous paragraph, the problem decomposes on a per-replenishment-period, per-location basis. We therefore wish to solve, for each t and r, the following problem:

$$v^*(x_{tr}) = \max \mathbb{E}\left[\sum_{j=1}^{d} \min\{Y_j, z_j\} \middle| X = x_{tr}\right], \qquad (2)$$

$$s.t. \sum_{j=1}^{d} z_j \le K_r, z_j \ge 0 \ \forall j = 1, \ldots, d,$$

where x_{tr} denotes auxiliary data available at the beginning of period t in the $(t, r)^{\text{th}}$ problem.

2.1. Internal Company Data

The internal company data collected consists of 4 years of sale and inventory records across the network of retailers, information about each of the locations, and information about each of the items. We aggregate the sales data by week (replenishment period of interest) for each feasible combination of location and item. We use these to collect data on Y,[1] and we include in X the sale volumes of each item at each location over each of the recent 3 weeks (as available; none for new releases), the total sale volume at each location over each of the recent 3 weeks, and the overall mean sale volume at each location over the past year. Information about retail locations includes to which chain a location belongs and the address of the location. We use the Google Geocoding API to parse the address and obtain precise coordinates of the location. We include in X indicators for the country and chain of the location. We also use coordinates to measure search attention as explained below. Information

about items include the medium (e.g., DVD or BluRay) and an item title. We disambiguate the item title to obtain a standardized title for the underlying content (e.g., movie name) and use this to collect information about the content as explained below.

Item Metadata, Box Office, and Reviews. To characterize the items and how desirable they may be to consumers, we harvest the data corresponding to each content title on IMDb.com and RottenTomatoes.com (RT). Using data from IMDb, we include in X the number of weeks since the original (e.g., theatrical) release data of the content, content type (film/TV), average user rating, number user ratings, number of awards (e.g., Oscars or Emmys) won and nominated, characteristic vector of first-billed actors' membership in 10 top communities (using Blondel et al. 2008) in the actor-movie graph, indicator vector of closest cluster in a hierarchical clustering of plot summaries by cosine similarity, characteristic vector of reported genres (out of 26), and MPAA rating (if rated). Using data from RT, we include in X the aggregate professional reviewers' score, average user rating, number user ratings, and American box office gross (for films). In Figure 1, we provide scatter plots and correlations of some of these attributes against sale figures in the first week of home entertainment (HE) release.

2.2. Search Engine Attention

To quantify the attention being given to different titles and to understand the local nature of such attention, we collect search query volume data from Google Trends (GT; www.google.com/trends).[2] GT provides data on the volume of Google searches for a given search term by time and geographic location. For each title, we measure the fraction of Google searches for the search term equal to the original content title in each week from 2011 to 2014 (inclusive) over the whole world, in each European country, and in each country subdivision (states in Germany, cantons in Switzerland, autonomous communities in Spain, etc.). We include in X the total search engine attention to each title over the first two weeks of original release globally, in the country, and in the country-subdivi-

Figure 1 Scatter Plots of Data from IMDb and RT (Horizontal Axes) against Total European Sales during First Week of HE Release (Vertical Axes, Rescaled) and Corresponding Coefficients of Correlation (ρ)

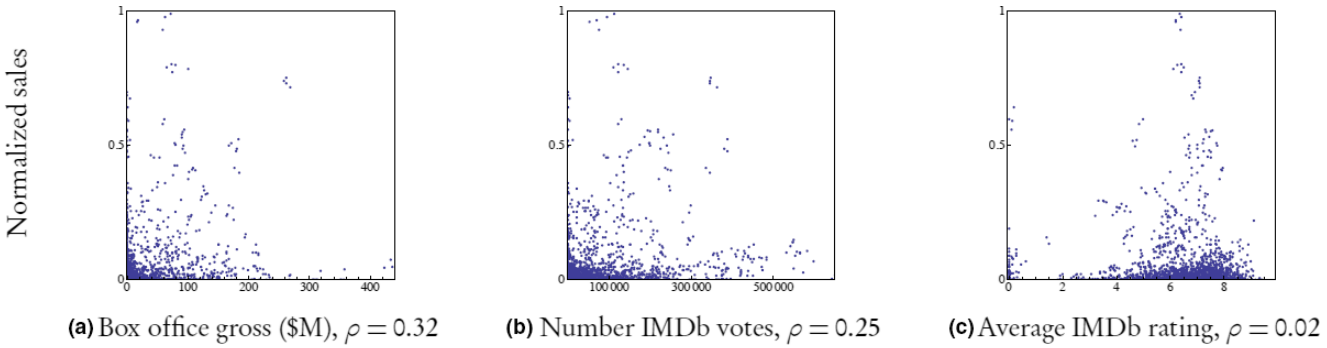
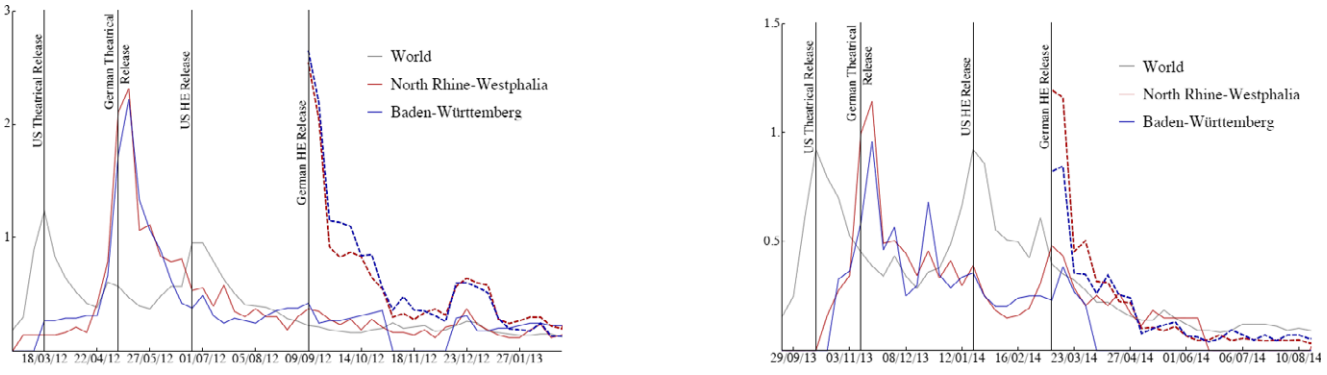
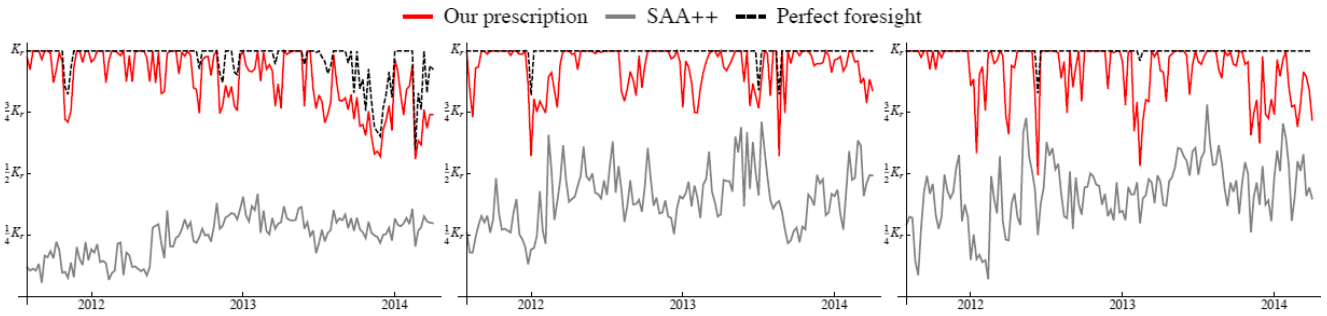


Figure 2 Weekly Search Engine Attention for Two Unnamed Films in the World and in Two German States (Solid Lines) and Weekly HE Sales for the Films in the Same STATES (Dashed Lines)



Note. Search engine attention and sales are both shown relative to corresponding overall totals in the respective region. The scales are arbitrary but common between regions and the two plots.

Figure 3 Sell-Through of Various Prescription Over Time



(a) Munich, Germany, $P = 0.89$ **(b)** Paris, France, $P = 0.90$ **(c)** The Hague, Netherlands, $P = 0.86$

sion of each location as well as the search engine attention to each title over each of the recent 3 weeks globally, in the country, and in the country-subdivision of each location. In Figure 2, we compare search engine attention to sales figures in two German states for two unnamed films, which shows the correlation of sales with local *local* search engine attention at original release and the ability of this attention to distinguish sale trends in two locations in the same country.

3. Inventory Prescriptions

Using the data described above, we construct inventory prescriptions $\hat{z}_N(x_{tr})$ for each location r and replenishment period t based on the local weighting approach based on random forest weights (see Bertsimas and Kallus 2014). To evaluate the prescription out-of-sample and as an actual live policy, we consider what we would have done over the 150 weeks from December 19, 2011 to November 9, 2014 (inclusive). At each week, we consider only data from time prior to that week to train the prescription and apply the prescription to the current week. Then, we observe what had actually materialized and score our performance. We compare the performance of our method with the performance of the perfect-forecast policy, which knows future demand exactly (no distributions) and the performance of a data-driven policy without access to the auxiliary data (SAA++).[3] When measured out-of-sample over the 150-week test period, we achieve a coefficient of prescriptiveness $P = 0.88$ averaged over the 20,000 locations, and, in Figure 3, we plot the performance over time at three specific locations. In other words, $P = 0.88$ means that our data X and our prescription $\hat{z}_N(x)$ gets us 88% of the way from the best data-poor decision to the impossible perfect-foresight decision in terms of sell-through volumes.

Notes

[1]In addressing problem (2) in a data-driven context, we face the issue that sales are a censored observation of demand Y. In Bertsimas and Kallus (2014), a remedy is provided in the form of a transformation based on a variant of the Kaplan-Meier method.

[2]Access to massive-scale querying and week-level trends data was generously provided by Google.

[3]For a fair comparison, because demand decay over product lifetime is significant, we let this policy depend on the distributions of product demand based on how long it's been on the market. Due to this handicap we term this policy SAA++.

References

Asur, S., B. Huberman. 2010. Predicting the future with social media. Proceedings of the IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology 492–499.

Bertsimas, D., N. Kallus. 2014. From predictive to prescriptive analytics. arXiv preprint arXiv:1402.5481.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, E. Lefebvre. 2008. Fast unfolding of communities in large networks. *J. Stat. Mech: Theory Exp.* **2008**(10): P10008.

Choi, H., H. Varian. 2012. Predicting the present with google trends. *Econ. Rec.* **88**(s1): 2–9.

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

Goel, S., J. Hofman, S. Lahaie, D. Pennock, D. Watts. 2010. Predicting consumer behavior with web search. *PNAS* **107**(41): 17486–17490.

Gruhl, D., L. Chavet, D. Gibson, J. Meyer, P. Pattanayak, A. Tomkins, J. Zien. 2004. How to build aWebFountain: An architecture for very large-scale text analytics. *IBM Syst. J.* **43**(1): 64–77.

Gruhl, D., R. Guha, R. Kumar, J. Novak, A. Tomkins. 2005. The predictive power of online chatter. Proceedings of the 11th ACM SIGKDD International Conference on Knowledge Discovery in Data Mining. 78–87.

Kallus, N. 2014. Predicting crowd behavior with big public data. Proceedings of the 23rd international conference on world wide web (WWW) companion. 625–630.