Integrated Forecasting and Inventory Control for Seasonal Demand: A Comparison with the Holt-Winters Approach

Gokhan Metan* Aurélie Thiele†

December 2007

Abstract

We present a data-driven forecasting technique with integrated inventory control for seasonal data and compare it to the traditional Holt-Winters algorithm. Results indicate that the data-driven approach achieves a 2-5% improvement in the average regret.

Keywords: data-driven algorithm, Holt-Winters algorithm, inventory management.

1 Introduction

Forecasting and optimization have traditionally been approached as two distinct, sequential components of inventory management: the random demand is first estimated using historical data, then this forecast (either a point forecast of the future demand or a forecast of the distribution) is used as input to the optimization module. In particular, the primary objective of time series analysis is to develop mathematical models that explain past data; these models are used in making forecasting decisions where the goal is to predict the next period' observation as precisely as possible. To achieve this goal, demand model parameters are estimated or a distribution is fitted to the data using a performance metric such as Mean Square Error, which penalizes overestimating and underestimating the demand equally. In practice, however, the optimization model penalizes under- and over-predictions unequally, e.g., in inventory problems backorder is viewed as particularly undesirable while holding inventory is more tolerated. In such a setting, the decision-maker places an order in each time period based on the demand prediction coming from the forecasting model, but the prediction of the forecasting model does not take into account the nature of the penalties in the optimization process and instead minimizes the (symmetric) error between the forecasts and the actual data points.

*Department of Industrial and Systems Engineering, Lehigh University, Bethlehem, PA 18015, gom204@lehigh.edu.
Work supported in part by NSF Grant DMI-0540143.
†Department of Industrial and Systems Engineering, Lehigh University, 200 W Packer Ave, Bethlehem, PA 18015, USA, aurelie.thiele@lehigh.edu. Corresponding author.

In this paper, we investigate the integration of the forecasting and inventory control decisions; in particular, our focus is on comparing the performance of this approach with the traditional Holt-Winters algorithm for random demand with a seasonal trend. The goal is not longer to predict future observations as accurately as possible using a problem-independent metric, but to blend the inventory control principles into the analysis to achieve superior inventory management. This work adds to the growing body of literature on data-driven inventory management (Godfrey and Powell 2001, Bertsimas and Thiele 2004, Levi et. al. 2006) by focusing on cyclical demand, and comparing the performance of a novel algorithm based on the clustering of data points (Metan and Thiele 2008) with that of the traditional Holt-Winters approach. Cluster creation and recombination allows the decision-maker to place his order at each time period based only on the most relevant data. To the best of our knowledge, we are the first authors to propose a clustering approach to the data-driven inventory management problem. As pointed out in Metan and Thiele (2008), customer behavior exhibits cyclical trends in many logistics applications where the influence of exogenous drivers is difficult to quantify accurately, which makes the approach particularly appealing. The proposed methodology captures the tradeoff between the various cost drivers and provides the decision-maker with the optimal order-up-to levels, rather than the projected demand. A cost decrease of 2-5% in experiments suggests that inventory managers could greatly benefit from implementing this approach.

The rest of the paper is organized as follows. In Section 2, we describe the methodology for integrated forecasting and inventory control in a data-driven framework. We present simulation results and compare the performance of the proposed approach with the traditional Holt-Winters algorithm in Section 3.

2 Integrated data-driven forecasting and inventory control

In this paper, our objective is to determine the optimal order for the newsvendor problem under cyclical demand using the integrated approach; however, the method can extended to other problem structures. The data-driven algorithm seeks to differentiate between the deterministic *seasonal effect* and the *stochastic variability* (see Figure 1) so that cycle stock and safety stock levels can be set accurately. We use the following notation:

Cost parameters:

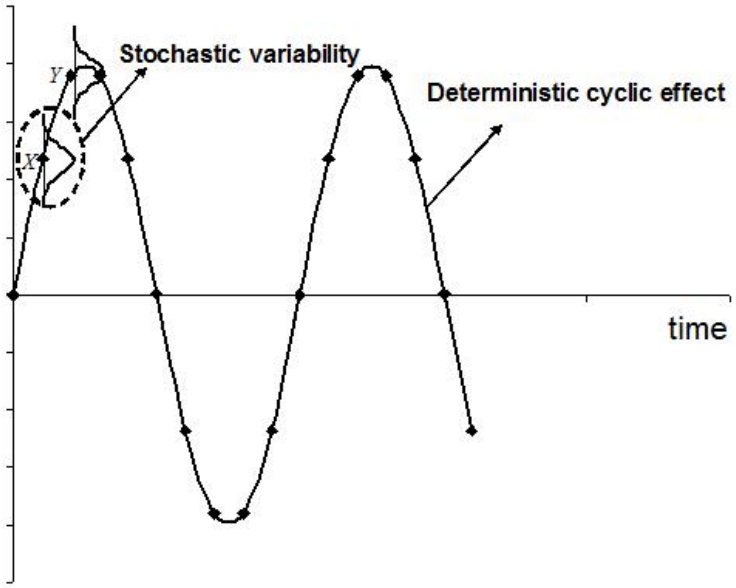

Figure 1: Differentiation of cyclic behavior and stochastic variability.

c: the unit ordering cost,

p: the unit selling price,

s: the salvage value,

c_u: undershoot cost ($c_u = p - c$),

c_o: overshoot cost ($c_o = c - s$),

α: critical ratio ($\alpha = \frac{c_u}{c_u + c_o}$).

Demand parameters and decision variable:

Q: the order quantity,

D: the demand level,

$\phi(.)$: pdf of standard normal distribution,

$\Phi(.)$: cdf of standard normal distribution,

S: the set of data points obtained from previous observations,

N: the number of historical observations in S,

d_i: the i^{th} demand observation in set S, with $i = 1, 2, \ldots, N$,

$d_{<i>}$: the i^{th} smallest demand in set S, with $i = 1, 2, \ldots, N$,

Additional parameters required for algorithm definition:

T: periodicity of time series,

C_j: list of data points assigned to cluster j, $j = 1, \ldots, T$, ranked in increasing order,

m: number of data points assigned to each cluster.

The methodology that we propose here is a *dynamic and data-driven approach* that builds directly upon the historical observations and addresses seasonality by creating and recombining clusters of past demand points. Our description follows Metan and Thiele (2008) closely, although at a less technical level, since our purpose here is not to demonstrate the algorithm's absolute performance but its relative performance with respect to the Holt-Winters method. *Clustering* can be defined as grouping a set of instances into a given number of subsets, called clusters, so that instances in the same cluster share similar characteristics. Most research efforts in clustering focus on developing efficient algorithms, in particular in the context of marketing applications and customer data (Jain et. al. 1999), which are beyond the scope of this paper. We only mention here the method of k-means, which was originally developed by MacQueen (1967) and is one of the simplest clustering approaches available in the literature: it partitions the data into k clusters by minimizing the sum of squared differences of objects to the centroid or mean of the cluster. The objective of minimizing total intra-cluster variability can be formulated as: $\min \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$, where there are k clusters S_i, $i = 1, 2, ..., k$ and μ_i is the mean point of all the instances x_j in S_i. This method presents similarities with the clustering approach we propose; in particular, we sort the demand and assign the data points to clusters in a manner that also attempts to minimize intra-cluster variability. A difference is that, while the k-means algorithm proceeds iteratively by updating the centroids of each cluster but keeps the same number of clusters throughout, we update the number of clusters by merging subsets as needed. Techniques for cluster aggregation/disaggregation are not discussed in this note; the reader is referred to Metan and Thiele (2008) for more details.

The master algorithm consists of five main steps, as described in Algorithm 2.1. For simplicity, we assume that the periodicity of the time series is known to the master algorithm; for settings where the true periodicity is unknown, an initial period estimation subroutine can be used to initiate the master algorithm.

Algorithm 2.1 (Master algorithm for cyclical demand (Metan and Thiele 2008))
Step 1. Sort the historical demand data in set S in ascending order to form the list $\widehat{S}$.
Step 2. Create T data clusters using $\widehat{S}$ (Algorithm 2.2).
***repeat** (until the end of the planning horizon)*
Step 3. Detect the phase φ and the corresponding cluster C_j of the demand at the next time period ($j = 1, 2, \ldots, T$).
Step 4. Select the next order level using inventory control policy π defined over the set C_j.
Step 5. Assign the new demand observation to appropriate cluster C_j ($j = 1, 2, \ldots, T$).
Update the phase for the next decision point ($\varphi \longleftarrow (\varphi + 1) \bmod T$)
end repeat.

Category and objective of each step of the master algorithm is summarized in Table 1. In Step

1 of Algorithm 2.1 the historical data is sorted and stored in the list $\widehat{S}$, which is then used by Algorithm 2.2 to create the initial clusters (Step 2). Algorithm 2.2 creates as many clusters as the periodicity of the seasonal demand function, i.e., one cluster for each phase of the seasonality, and assigns $m = \left\lfloor \frac{N}{T} \right\rfloor$ data points to each cluster, with the $N - mT$ oldest data points being discarded. Alternatively, the last cluster could receive more points than the others; for the sake of clarity, we present here the version where all the clusters have the same number of points.

Table 1: Categorization of master algorithm's steps.

Activity Category	Objective	Steps
Forecasting	Initialize forecasting method	Steps 1 & 2
	Forecast cyclic effect	Step 3
	Update forecasts	Step 5
Inventory control	Make optimal ordering decision given the forecast	Step 4

The output of Step 2 is a set of clusters $\{C_j | j = 1, 2, \ldots, T\}$ and each cluster can be thought of as a set that defines a group of possible forecasts corresponding to a given phase of the time series. The algorithm selects such a set C_j among the available sets $\{C_j | j = 1, 2, \ldots, T\}$ in Step 3, which is equivalent to defining a set of predictions rather than having a single point forecast. Then, Step 4 concludes the decision-making process by selecting the order-up-to level using inventory control policy $\pi : C_j \rightarrow \Re$. Finally, Step 5 updates cluster C_j by assigning the new observation to this set and the process repeats itself starting from Step 3.

Algorithm 2.2 (Initial clustering) *Set* $m = \left\lfloor \frac{N}{T} \right\rfloor$, $j = 1$.
repeat
Assign observations $d_{<m(j-1)+1>}, \ldots, d_{<m(j-1)+m>}$ *to cluster* C_j,
$j \longleftarrow j + 1$,
end repeat *when* $j = T + 1$.

Figure 2 shows the repartition of 100 data points into clusters for a cyclical demand with a periodicity of $T = 20$. Since, $m = \frac{N}{T} = \frac{100}{20} = 5$, the algorithm creates 20 clusters with 5 data points in each. The data points that lie within the boundaries of two consecutive horizontal lines in Figure 2 all belong to the same cluster.

The newsvendor problem, which is the focus of this work, has been extensively studied in the literature under a wide range of assumptions (see, e.g., Porteus 2002.) We review here properties of the optimal order when the demand D is a continuous non-negative random variable with probability density function (pdf) f and cumulative density function (cdf) F, and there is no setup cost for placing an order. The newsvendor decides how many items to order before knowing the

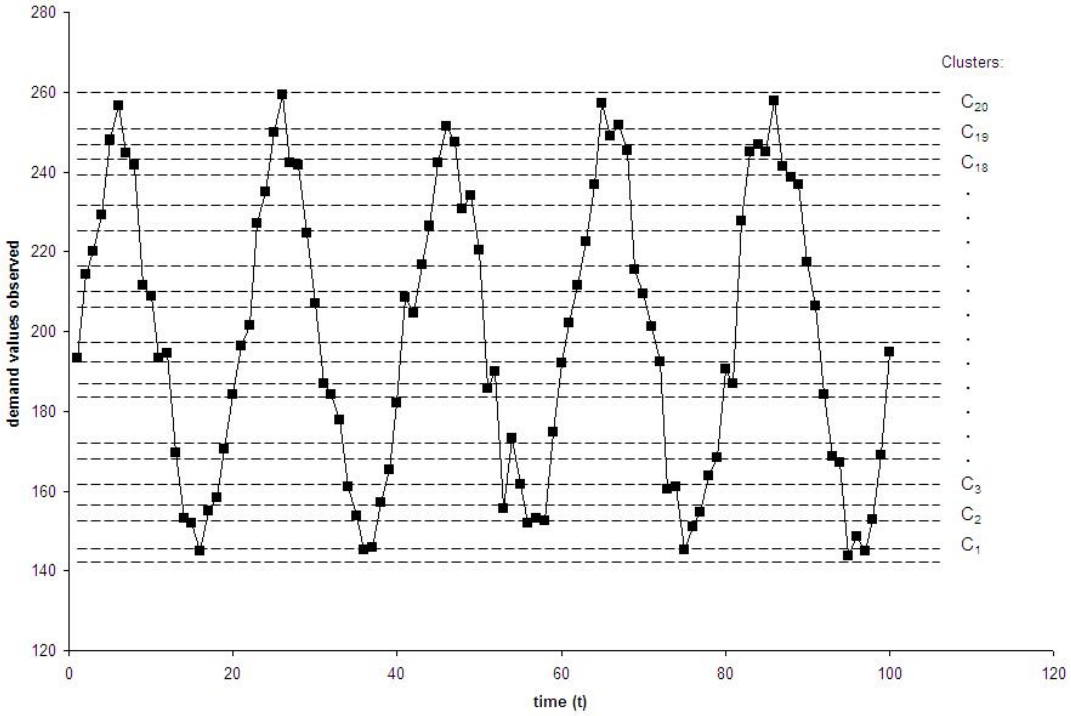

Figure 2: Clustering the historical data for $D \sim N(200 + 50\sin(\frac{2\pi t}{20}), 7)$. (Metan and Thiele 2008)

exact value of the demand, with the goal of maximizing his expected profit. Since the goods are perishable, no inventory can be carried over to the next time period. The classical newsvendor problem is then formulated as:

$$\max_{Q \geq 0} (p - c)\, Q - (c - s)\, E\left[\max(0, Q - D)\right],$$

The optimal solution to the classical newsvendor problem is (Porteus 2002):

$$Q^* = F^{-1}(\alpha)$$

The value of the optimal order depends heavily on the underlying demand distribution. In practice, however, such precise knowledge is difficult to obtain, because we lack historical observations to compute meaningful estimates. This data scarcity may be due to the introduction of a new product (no historical data is available), or to the non-stationarity of the demand (historical data is then in large part irrelevant, since past demands obey different distributions than the demand at the current time period.) This motivates the *direct use* of the empirical data to guide the manager, so that the decision-making process builds upon the precise amount of information available without estimating any additional parameters.

In what follows, we focus on non-stationary demand as the main reason for scarce data. We assume that we do not have any information about the distribution of the underlying demand process, but we do have a set of historical observations at our disposal. The data-driven counterpart of the classical newsvendor problem becomes:

$$\max_{Q \geq 0} (p - c)\, Q - \frac{(c - s)}{N} \sum_{i=1}^{N} \max(0, Q - d_i),$$

and the optimal order Q^* is given by:

$$Q^* = d_{<j>}, \text{ with } j = \lceil \alpha N \rceil.$$

where all d_i are possible realizations of the demand at the next time period.

In the problem with seasonal demand, we define the control policy as the α-quantile of the data set C_j, rather than the whole data set (which contains data points observed for different phases of the cycle). Therefore, we restate Step 4 of the master algorithm as follows:

Step 4. Select the next order level using inventory control policy $\pi : Q^* = d_{<\lceil \alpha |C_j| \rceil>}$ *defined over the set* C_j.

3 Experimental Results

3.1 The Holt-Winters Method

We start by reviewing the traditional procedure. The Holt-Winters method is a type of exponential smoothing technique that is best suited for forecasting a time series with a linear trend and seasonality. The technique is due to Winters (1960) and the method assumes that the time series can be described by the model:

$$y_t = (\beta_0 + \beta_1 t) SN_t + \epsilon_t,$$

where β_0, β_1, SN_t are the *permanent component* (or intercept), *trend*, and *seasonal factor*, respectively. These parameters are estimated using the historically available data and the estimates are calculated using Equations (1)-(3). Let us assume that we are at the end of period $T-1$ and already have the estimates of the parameters β_0, β_1, SN_t (the estimates of the parameters at time $T-1$ are denoted as $a_0(T-1)$, $b_1(T-1)$, and $sn_T(T-L-1)$, respectively). In period T, we observe a new data point, y_T, and we would like to update our estimates using this new information. By using Equations (1)-(3), we can obtain the new estimates of the model (i.e., $a_0(T)$, $b_1(T)$, and $sn_T(T)$) and forecast next period's value via Equation (4). In these equations L is the periodicity of the process and α, β, γ are the smoothing constants. In Equation 4, τ is the variable that defines the

number of periods from the current period in which the forecast is being made. For instance, if we are at period $T = 10$ and want to forecast the time series' value at time 12, then $\tau = 2$ and the corresponding prediction is denoted by $\hat{y}_{T+\tau}(T)$.

$$
\begin{aligned}
a_0(T) &= \alpha \frac{y_T}{sn_T(T-L)} + (1-\alpha)\left[a_0(T-1) + b_1(T-1)\right] && (1)\\
b_1(T) &= \beta\left[a_0(T) - a_0(T-1)\right] + (1-\beta) b_1(T-1) && (2)\\
sn_T(T) &= \gamma \frac{y_T}{a_0(T)} + (1-\gamma) sn_T(T-L) && (3)\\
\hat{y}_{T+\tau}(T) &= \left[a_0(T) + b_1(T)\tau\right] sn_{T+\tau}(T+\tau-L) && (4)
\end{aligned}
$$

One of the difficulties with the Holt-Winters method is setting the optimal values of the smoothing constants, α, β, and γ. These constants take their values in [0, 1]. An appropriate combination of smoothing constants is found by minimizing a performance metric, such as Mean Square Error (MSE), over the historical data set.

3.2 Results

In this section, we implement the Holt-Winters method and the integrated data-driven forecasting technique proposed in Section 2. We compare the performances of both techniques in terms of the average regret via simulation.

The underlying demand process used in the experiments is modeled as a Normally distributed demand, $N(\mu, \sigma^2)$, where $\mu = a + b\sin(\frac{2\pi}{T})$. We consider two sets of experiments as shown in Table 2. In Set 1 and Set 2 of experiments we consider high and moderate levels of amplitude for the seasonality (i.e., $b = 200$ as high and $b = 50$ as moderate). We use the value of 500 for the demand level (i.e., $a = 500$) and standard deviation of $\sigma = 7$. At the beginning of the simulation, we generate demand observations from the underlying demand distribution for a length of time defined by the warm-up period. This set of demand observations is used by the Holt-Winters method and the data-driven method for initializing their own parameters. We set the optimal values for the smoothing constants (i.e., α, β, γ) of the Holt-Winters method with 0.05 precision. For each experimental setting we get 20 simulation replications with each replication having a run length of 5000 time periods, and collect the performance statistics of each method. We use the following values for the inventory cost parameters: $p = 10$ (the unit selling price); $c = 7$ (the unit purchasing cost), and $s = 5$ (the unit salvage value). Finally, there are different methods that can be used to initialize the Holt-Winters method's starting estimates. Here, we use the one described in Bowerman and O'Connell (1979).

When we examine the results of individual simulation runs, we see that the data-driven method always outperforms the Holt-Winters method in terms of average regret values. To further strengthen

Table 2: Experimental Parameters.

Exp. Set	b	T	σ	Warm-up	Simulation length
Set 1	200	10	7	200	5000
	200	20	7	200	5000
	200	100	7	1000	5000
Set 2	50	10	7	200	5000
	50	20	7	200	5000
	50	100	7	1000	5000

our analysis, we construct the 95% paired-t confidence intervals. Table 3 summarizes the results including the 95% confidence intervals for each experiment. None of the confidence intervals given in Table 3 contains the value 0, which allows us to conclude that the average regret achieved by using the data-driven method is significantly less than the average regret achieved by Holt-Winters method with 0.95 confidence level. Also, experiments indicate that the data-driven method provides 2-5% improvement in average regret over the Holt-Winters method for the experimental setting tested. The improvement is greater in experiments with large amplitude.

Table 3: Summary of results.

Experiment	Avg. Performance of Data-Driven meth.	Avg. Performance of Holt-Winters meth.	Avg. Performance Difference	95% CI
Set 1, Exp. 1	13.7645	14.418	0.6535	[0.5712, 0.7357]
Set 1, Exp. 2	13.6365	14.3975	0.761	[0.6949, 0.8270]
Set 1, Exp. 3	13.762	14.4255	0.6635	[0.6348, 0.6921]
Set 2, Exp. 1	13.739	14.29	0.551	[0.4981, 0.6038]
Set 2, Exp. 2	13.6335	14.295	0.6615	[0.6136, 0.7093]
Set 2, Exp. 3	14.1155	14.423	0.3075	[0.2124, 0.4026]

We also investigate the impact of the length of the planning horizon on the performances of the methods. Figure 3 shows the average simulated regret values for both methods as a function of simulation run length. From the figure, we can see that the performance of both methods improve as time elapses and the rate of improvement is higher early in the process as compared to the rate of improvement near the end of the planning horizon. Also, the performances of both methods stabilize once enough time periods have elapsed.

Our preliminary simulation results indicate that the integrated data-driven forecasting and inventory control technique improves the performance of the system over the Holt-Winters method. Now, we perform additional experiments to answer the following questions: (i) What are the performances of both methods measured in terms of mean square error (MSE) as well as average

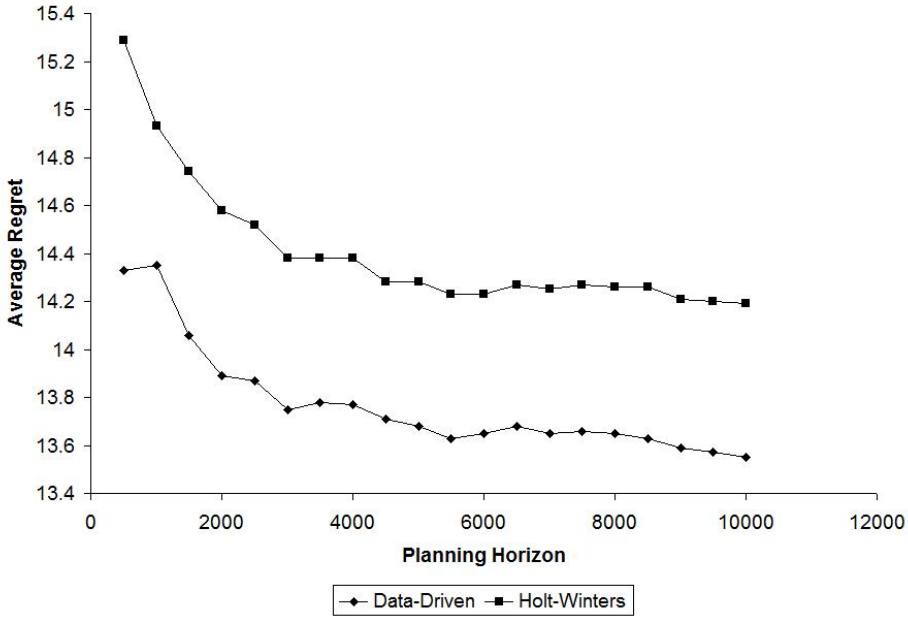

Figure 3: Effect of planning horizon on the methods' performance (results obtained under Set 1, Experiment 3 parameter levels).

regret? (ii) What impact does a different inventory control policy $\widehat{\pi}$ have on the performance of the data-driven technique? (iii) What is the impact of the cost parameters on the relative performance of the forecasting techniques?

We investigate these issues under the same experimental conditions as above, excepted for the number of simulation replications (we take 60 replications rather than 20 to obtain tighter confidence bounds), by extending the first experimental set in Table 2. The new experimental instances are shown in Table 4, in which the first three instances impose a critical ratio of $\alpha = \frac{c_u}{c_u+c_o} = 0.6$ whereas the second three instances have a critical ratio of $\alpha = \frac{c_u}{c_u+c_o} \approx 0.78$. Also, to test the impact of inventory control policy on the system performance, we use the following alternative policy in Step 4 of the master algorithm.

Step 4. Select the next order level using inventory control policy $\widehat{\pi} : Q = \frac{\sum_{d_i \in C_j} d_i}{|C_j|}$ *defined over the set* C_j.

The reason for using the cluster averages as the order-up-to level policy is that it minimizes the mean-square error (MSE) and it provides neutral predictions against under- and over-shooting the demand. In other words, it provides forecasts similar to those of the Holt-Winters method since it now discards the imbalance between the cost parameters when producing the predictions.

Results of the second set of experiments are summarized in Tables 5 and 6 (confidence intervals in Table 6 are calculated for the average regret difference between the two methods). In terms

Table 4: Experimental Parameters.

Instance	b	T	σ	Warm-up	Simulation length	p	c	s
I1	200	10	7	200	5000	10	7	5
I2	200	20	7	200	5000	10	7	5
I3	200	100	7	1000	5000	10	7	5
I4	200	10	7	200	5000	10	3	1
I5	200	20	7	200	5000	10	3	1
I6	200	100	7	1000	5000	10	3	1

of average regret, the data-driven method (under both π and $\widehat{\pi}$ policies) performs significantly better than the Holt-Winters method. Also, the proposed method performs even better when π is used as the inventory control policy, which is the optimal policy for the newsvendor problem. This indicates that under the problem-specific optimal control policy, integrated forecasting and inventory control results in significantly improved performance. In terms of MSE, the Holt-Winters method results in lower values than the data-driven method operating under the policy π, which matches the intuition. However, when the policy is changed to $\widehat{\pi}$, MSE values of the data-driven method becomes slightly lower than the Holt-Winters method; this indicates that the data-driven approach also has the power of closely predicting the actual observations when an appropriate policy such as $\widehat{\pi}$ is used.

Table 5: Experimental Results.

Instance	Holt-Winters		Data-driven with π		Data-driven with $\widehat{\pi}$	
	Avg. Regret	MSE	Avg. Regret	MSE	Avg. Regret	MSE
I1	14.4127	52.0754	13.7557	53.7847	14.1444	50.2954
I2	14.4316	52.2010	13.5916	52.8007	14.0038	49.3104
I3	14.4611	52.5743	13.8555	52.7820	14.2142	50.6118
I4	25.9762	52.0754	19.1052	78.4375	25.4281	50.2954
I5	25.9990	52.2010	18.8668	78.0998	25.1729	49.3104
I6	26.0188	52.5610	19.7412	67.5530	25.7455	50.7539

Another performance statistics is presented in Table 7. As discussed above, the imbalance between the cost parameters is not taken into account when producing the forecasts in the Holt-Winters method. Utilizing this information is key in the integrated forecasting and inventory control. When the imbalance between the cost parameters increases, the percentage of improvement in the average regret increases. To quantify the impact of the critical ratio on the average regret, we perform additional experiments for instance I1 for different cost parameter combinations. Figure 4 shows the percentage of improvement in the average regret values with respect to the critical

ratio. We calculate the percentage of improvement by comparing the average regret values of the data-driven (under policy π) and Holt-Winters methods. The percentage of improvement is a convex function of the critical ratio and reaches the minimum value at 0.5, which is the point where under- and over-shooting the demand are penalized equally. Therefore, the integrated forecasting and inventory control loses its attractiveness since bare forecasting achieves the same performance. However, when the imbalance is high (in many practical applications, backorder costs are much higher than holding costs), the cost reduction becomes very significant: up to 90% cost reduction is observed at extreme critical ratio values (Figure 4).

Table 6: Experimental Results.

Instance	95% CI for the difference between Holt-Winters and Data-driven with π		95% CI for the difference between Holt-Winters and Data-driven with $\widehat{\pi}$	
	Lower bound	Upper bound	Lower bound	Upper bound
I1	0.6128	0.7012	0.2243	0.3123
I2	0.8027	0.8773	0.3860	0.4695
I3	0.4966	0.7146	0.0920	0.4018
I4	6.6637	7.0784	0.3406	0.7556
I5	6.9473	7.3172	0.6275	1.0247
I6	6.1075	6.4478	-0.3252	0.8718

Table 7: Percentage of improvement in average regret.

$\alpha = 0.6$			$\alpha \approx 0.78$		
I1	I2	I3	I4	I5	I6
4.56%	5.82%	4.19%	26.45%	27.43%	24.13%

Acknowledgments

The work of the first author is supported in part by NSF Grant DMI-0540143. The work of the second author is supported in part by NSF Grant DMI-0540143 and an IBM Faculty Award.

References

Bertsimas, Dimitris, and Aurélie Thiele. (2004). A data-driven approach to newsvendor problems. Technical report, Massachusetts Institute of Technology, Cambridge, MA.

Bowerman B. L., and Richard T. O'Connell. (1979) *Forecasting & Time Series.* Duxbury Press, Belmont, CA.

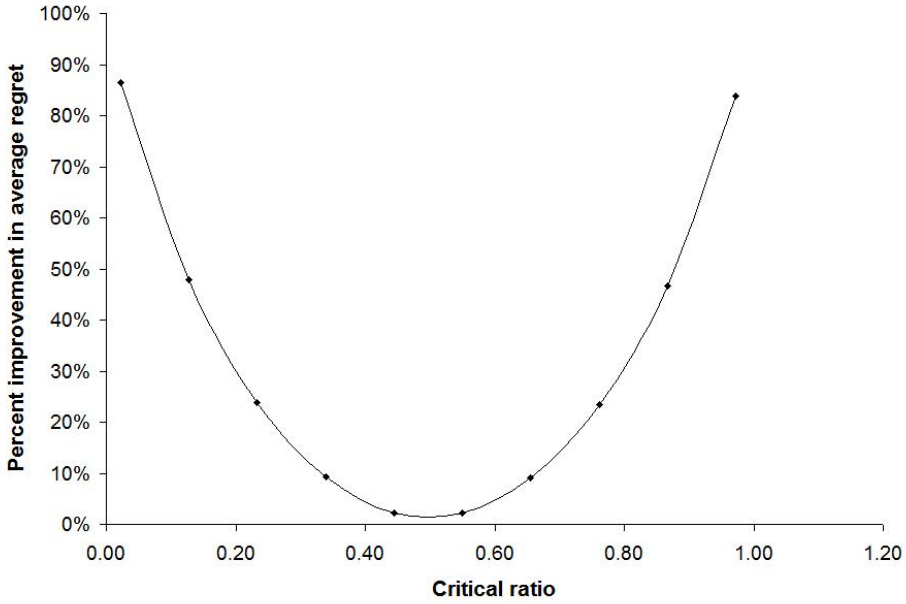

Figure 4: Effect of imbalance between cost parameters on the percentage of improvement in average regret (results obtained using the parameter levels of instance I1).

Godfrey, Gregory, and Warren Powell. (2001). An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science,* **47** 1101-1112.

Jain, A. K., and M. N. Murty, and P. J. Flynn (1999) Data Clustering: A Review. *ACM Computing Surveys,* **31** 264-323.

Levi, Retsef, Robin Roundy, and David Shmoys. (2006). Provably near-optimal sampling-based policies for stochastic inventory control models. *Proceedings of the 38th annual ACM Symposium on the Theory of Computing (STOC).*

MacQueen, J. B. (1967) Some Methods for Classification and Analysis of Multivariate Observations. *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability,* Berkeley, University of California Press, **31** 264-323.

Metan G., A. Thiele. (2008) A dynamic and data-driven approach to the newsvendor problem under seasonal demand. Computational Optimization and Logistics Challenges in the Enterprise. Book chapter, ed. Springer, to appear.

Porteus, E. (2002) *Stochastic Inventory Theory*, Stanford University Press, Palo Alto, CA.

Winters P. R (1960) Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, **6** 324-342.