Proceedings of the ASME 2018 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference IDETC/CIE 2018 August 26-29, 2018, Quebec City, Quebec, Canada

DETC2018-85710

PREDICTING PURCHASE ORDERS DELIVERY TIMES USING REGRESSION MODELS WITH DIMENSION REDUCTION

Jundi Liu
Industrial & Systems Engineering
University of Washington
Seattle, WA, 98195

Steven Hwang
Industrial & Systems Engineering
University of Washington
Seattle, WA, 98195

Walter Yund
General Electric Global Research
Niskayuna, NY, 12309

Linda Ng Boyle
Industrial & Systems Engineering
Civil & Environmental Engineering
University of Washington
Seattle, WA, 98195

Ashis G. Banerjee
Industrial & Systems Engineering
Mechanical Engineering
University of Washington
Seattle, WA, 98195

ABSTRACT

In current supply chain operations, the transactions among suppliers and original equipment manufacturers (OEMs) are sometimes inefficient and unreliable due to limited information exchange and lack of knowledge about the supplier capabilities. For the OEMs, majority of downstream operations are sequential, requiring the availabilities of all the parts on time to ensure successful executions of production schedules. Therefore, accurate prediction of the delivery times of purchase orders (POs) is critical to satisfying these requirements. However, such prediction is challenging due to the suppliers' distributed locations, time-varying capabilities and capacities, and unexpected changes in raw materials procurements. We address some of these challenges by developing supervised machine learning models in the form of Random Forests and Quantile Regression Forests that are trained on historical PO transactional data. Further, given the fact that many predictors are categorical variables, we apply a dimension reduction method to identify the most influential category levels. Results on real-world OEM data show effective performance with substantially lower prediction errors than supplier-provided delivery time estimates.

INTRODUCTION

Traditional supplier-buyer transactions suffer from serious shortcomings due to restricted communication among the entities and inaccurate knowledge of parts availabilities, leading to large variations in production lead times and frequent downstream schedule changes. The conventional solution to this issue is for the original equipment manufacturers (OEMs) to require their suppliers, who are usually small and medium-scale enterprises (SMEs), to share their production schedules and provide the estimated delivery times of the purchase orders (POs). While the first requirement is nearly impossible for the SMEs to adhere in order to maintain competitive advantages, the second requirement is difficult to carry out in practice. The SMEs production plans depend on the timely deliveries of raw materials from other lower tier suppliers as well as the fluctuating, and sometimes urgently placed, demands for the produced parts. Moreover, the OEMs procure parts, many of which have lead times of several weeks to more than a year, from thousands of distributed suppliers around the world. Each supplier has different capability and capacity to provide parts of various types, quantities, and qualities. All these characteristics render the problem of estimating the delivery time of a specific PO extremely challenging.

From the OEMs perspectives, accurately predicting the availabilities of the purchased parts are essential for adequate stocking of inventories and maintenance of uninterrupted assembly operations. The main uncertainty comes from the complete reliance on the promise (estimated delivery) dates provided by the suppliers. For example, the suppliers may update the promise dates continuously, which may cause the OEMs to adjust their production plans accordingly. In other situations, the promised delivery dates or the latest status of the POs may never be updated, which may result in a lack of communication on the anticipated problems of delivering the orders. In general, the OEMs are not able to control the delivery times of their POs. To be more specific, when the promised delivery dates are relatively short, they can be treated as a significant predictor of the actual delivery dates. However, if the promised delivery dates are very long, the estimation uncertainty increases drastically, and they become a weak predictor of the actual delivery dates, making them untrustworthy for the OEMs. Therefore, not only the OEMs, but the suppliers need more accurate prediction models to estimate the delivery times of the POs, and, thereby, increase the overall level of trust in the supply chain.

We address these challenges by introducing two supervised machine learning models, namely, Random Forest and Quantile Regression Forest. These models learn from the historical transactional data on closed POs. Subsequently, the models can be used to predict the delivery times of future open POs. They consider the essential information of the POs as predictor variables, which are either continuous (e.g., quantity ordered, estimated delivery time) or categorical (e.g., item type, part flow indicator). However, the number of categorical levels that can be considered are limited for both these models. Hence, we augment the models with additional categorical predictors using a principal components analysis (PCA)-based dimension reduction method. The augmented models are then applied to the historical data to identify additional significant predictors related to the suppliers and ordered parts.

Test results on actual closed POs demonstrate the effectiveness of our regression models and the dimension reduction method. Prediction errors are substantially smaller than those for the supplier estimated delivery dates. The prediction errors decrease even more after applying the dimension reduction method. By implementing these prediction models, the OEMs and suppliers can use their distinct historical data to infer more accurate estimates of the delivery times for their open POs. As a result, they would be able to adapt their production schedules and inventory management strategies to ensure fulfillment of the customer demands.

RELATED WORK

Our work is related to supply chain forecasting using advanced machine learning models and dimension reduction methods. In today's inter-connected economy, supply chain forecasting has gained a lot of attention. One similar work is based on a hybrid statistical method for aircraft engine parts delivery times prediction [1]. However, categorical variables are not considered in that work. To the best of our knowledge, no specific work has been done on predicting PO delivery times using Random Forests or Quantile Regression Forests. For dimension reduction, there have been plenty of well-defined methods such as PCA, variants of PCA, such as Kernel PCA and Graph-based Kernel PCA, Linear or Generalized Discriminant Analysis (LDA/GDA), and Non-negative Matrix Factorization (NMF) [2]. While these methods have been shown to be effective for a variety of real-world problems, we are not aware of their use in supply chain forecasting.

Most supply chain studies have, instead, focused on supplier delivery performance measurement [3], supply chain optimization [4], and manufacturing suppliers classification. Recent use of text mining through semantic modeling and Naïve Bayes classification to categorize the suppliers based on their capabilities are shown in [5,6]. Another work to improve the communication among the suppliers and buyers using a matchmaking algorithm is provided in [7].

As regards supply chain performance enhancement, one of the earliest works is seen in [8]. A subsequent analysis on the impact of model selection for demand forecasting, inventory replenishment decisions by the retailers, and production decisions by the suppliers is demonstrated in [9]. Demand forecasting and order lead times have been investigated using a simple forecasting model in [10]. Autoregressive model based multi-step demand forecasting is used to improve inventory management performance in [11]. A hybrid intelligent system combining Autoregressive Integrated Moving Average (ARIMA) models and neural networks is shown to improve forecasting accuracy in [12]. Comparisons of advanced machine learning methods such as neural networks, recurrent neural networks, and support vector machines with traditional linear forecasting models are provided in [13]. And, a quantile regression based case study for biomass supply chain optimization under uncertainty is found in [14].

Collaborative forecasting and planning in supply chain systems have been widely studied in recent years. An example of a Japanese manufacturer has been investigated in the context of internal and external collaborations in [15]. The effects of collaborative planning, forecasting, and replenishment in supply chains are examined in [16]. The impact of supply chain collaboration on organizational performance has been studied in [17]. However, in any real-world supply chain system, only partial collaboration is feasible via limited sharing of information among the suppliers and OEMs. For example, the suppliers might provide initial estimates of the PO delivery times without any information about their core capabilities. Our study falls within this paradigm of partial collaboration, which has not been explored much so far.

TECHNICAL METHOD

In this section, we present our prediction models as the following two-step method: a) supervised learning on historical POs transactions data to predict the expected means and quantiles of their delivery dates; and b) dimension reduction to select the dominant levels of the categorical variables in the transactional data set to enhance prediction performance.

Random Forest and Quantile Regression Forest

Random Forest [18] and Quantile Regression Forest [19] are both ensemble learning methods for supervised learning. They are constructed by a multitude of decision trees at the training time, which aim to robustly capture the underlying nonlinear trends and consider different measures of central tendency and statistical dispersion. Furthermore, they usually provide accurately predictions of unobserved data, prevent overfitting to training samples, and avoid undue influence of the outliers in the datasets. Therefore, it is not surprising that they have found a lot of success in diverse application domains, including banking [20], healthcare [21], and e-commerce [22]. The difference between the two methods lies in their outputs: for Random Forest, it is the mean prediction, whereas, for Quantile Regression Forest, it is the quantile prediction of the individual trees. To present both the methods used in our work, we begin with the underlying decision trees.

There are two types of decision trees, classification tree and regression tree, of which the latter one is of interest here. Regression tree model uses a tree structure to partition the dataset recursively and compute specific regression values at the leaf node for different conjunctions of predictor values. The regression tree model uses variance reduction to identify the most suitable for splitting any interior node into two branches. Variance reduction, $I_V(N)$, computes the total reduction of the variance of the target variable x due the split of a node N as shown in Eqn. (1).

$$I_V(N) = \frac{1}{|S|^2}\sum_{i\in S}\sum_{j\in S}\frac{1}{2}(x_i-x_j)^2 - \left(\frac{1}{|S_t|^2}\sum_{i\in S_t}\sum_{j\in S_t}\frac{1}{2}(x_i-x_j)^2 + \frac{1}{|S_f|^2}\sum_{i\in S_f}\sum_{j\in S_f}\frac{1}{2}(x_i-x_j)^2\right) \quad (1)$$

where S, S_t, and S_f are the set of presplit sample indices, set of sample indices for which the split test is true, and set of sample indices for which the split test is false, respectively. x_i and x_j are the samples within the set of sample indices for computing the variance. The above summands are, indeed, variances for each set of the data. Thus, Eqn. (1) calculates the variance reduction in each split.

By recursively computing the variance reduction for each split, the learning goal is successfully achieved by identify the splitting variable with the optimal split of the dataset. Therefore, the regression tree model is gradually expanded until there is no variance reduction in any split.

Given the regression tree model, a random forest model is built on it consisting of multiple simple regression tree models. The terminology "Random" often refers to two sources: randomly sampled training dataset by applying Bootstrap [23] on the original dataset, and randomly selected subset of features to choose the best split in each tree model. It is a simple but effective mechanism to aggregate many simple models to tackle a complex prediction task.

The practical training algorithm for random forest model applies the general technique of Bootstrap. Given a training dataset $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ with responses $\mathbf{Y} = (y_1, \ldots, y_n)$, Bootstrap repeats the following steps M times:

1. For $m = 1, \ldots, M$, sample n training examples with replacement from $\mathbf{X}, \mathbf{Y}$. Call them $\mathbf{X}_m$, $\mathbf{Y}_m$.
2. Train a regression tree f_m on $\mathbf{X}_m$, $\mathbf{Y}_m$.

After training a desired number of regression trees, prediction for a new sample $\mathbf{x}'$ can be made by taking the average prediction value of all individual regression trees using Eqn. (2).

$$\hat{y} = \frac{1}{M}\sum_{m=1}^{M} y_m(\mathbf{x}') \quad . \quad (2)$$

The prediction of a single regression tree may be highly sensitive to noise and outliers, but the average of many trees is not, as long as the trees are not strongly correlated. Bootstrap is the way of preventing correlations in the trees by randomly sampling with replacements from different training sets. However, the practical implementation of random forest model is not exactly the same as the above procedure. In order to gain better computational and learning performance, each tree is trained using only a subset of the predictors in the training dataset. This process is another way of confirming that the trees are uncorrelated. Since the predictors are not guaranteed to be independent, using the highly correlated features to train the regressing tree will cause serious correlation issues even using bootstrap samples. Therefore, training with randomly sampled predictors ensures the uncorrelation property of the trees.

As a generalization of Random Forest, Quantile Regression Forest gives the predictions of the conditional quantiles instead of the conditional mean. In statistical analysis, the goal is to infer the relationship between the response variable and the predictors. The standard regression analysis is trying to develop an estimate of the conditional mean of the response variable. The conditional mean minimizes the expected squared error loss in Eqn. (3).

$$E(Y|X=x) = \arg\min_z E\{(Y-z)^2|X=x\} \quad . \quad (3)$$

However, beyond the conditional mean, the quantile regression analysis better describes the conditional distribution of the response. For a continuous distribution function, the α quantile $Q_\alpha(x)$ is defined in Eqn. (4).

$$Q_\alpha(x) = \inf\{y : F(y|X=x) \geq \alpha\} \tag{4}$$

where $F(y|X=x) = P(Y \leq y|X=x)$ is the conditional distribution function of Y.

Quantile regression forest aims to estimate the conditional quantiles instead of conditional mean from data. Like minimizing the squared error loss function, it uses the weighted absolute deviations defined in Eqn. (5) as the loss function.

$$L_\alpha(y,q) = \begin{cases} \alpha|y-q| & y > q \\ (1-\alpha)|y-q| & y \leq q \end{cases} \tag{5}$$

where α is the desired quantile, and q is a constant number.

And the training process is achieved by minimizing the expectation of the weighted absolute deviations defined in Eqn. (6).

$$Q_\alpha(x) = \arg\min_q E\{L_\alpha(Y,q)|X=x\} \quad . \tag{6}$$

The above optimization problem can be solved efficiently due to the convex property using a parametric method [24], or a non-parametric approach [25] [26].

After the estimation of the parameters, the prediction of the quantile regression forest is done in two steps. First, estimate the conditional distribution of the response variable, and then estimate the conditional quantiles. The conditional distribution function of Y given $X=x$ is defined as

$$F(y|X=x) = P(Y \leq y|X=x) = E(\mathbf{1}_{\{Y \leq y\}}|X=x) \tag{7}$$

where $\mathbf{1}$ is the indicator function, and the last expression has the form of conditional expectation. Therefore, we employ a method, similar to that for Random Forest, to estimate the conditional distribution. Using Eqn. (2), the estimate of the distribution is given by

$$\hat{F}(y|X=x) = \sum_{i=1}^{n} \mathbf{1}_{\{\hat{y} \leq y\}} \quad . \tag{8}$$

Using the similar procedure of Random Forest training, an estimate $\hat{F}(y|X=x)$ is obtained. Finally, by plugging $\hat{F}(y|X=x)$ into Eqn. (4), we get the estimate $\hat{Q}_\alpha(x)$ of the conditional quantile $Q_\alpha(x)$. After successfully training the Quantile Regression Forest, it is used to predict any quantile of the condition distribution of the response variable.

Dimension Reduction

In machine learning and statistical regression analysis, categorical predictors with a large number of levels are a challenge. By default, the algorithms will expand the categorical predictors into dummy variables in data preprocessing. For categorical predictors with many levels, the algorithms are likely to run into the problem often called curse of dimensionality [27].

PCA was developed in early 20th century [28], and is a well-known technique for dimension reduction. Basically, it uses an orthogonal transformation to convert a set of possibly correlated predictors into a set of linearly uncorrelated predictors called principal component. PCA is implemented by Singular Value Decomposition (SVD) of the design matrix or Eigenvalue Decomposition (EVD) of the covariance matrix. There are several effective methods and applications of applying PCA on discrete data [29].

Our dimension reduction process, i.e. level selection for categorical predictors, is relying on the fundamental of PCA. However, given the computational complexity of the matrix decomposition for categorical predictors with too many levels, it is necessary to pre-process the data. In summary, the dimension reduction is implemented via the following steps:

1. Keep the levels that maintain 80% of the total rows, and combine the remaining levels into a single level.
2. Expand the categorical predictors into dummy variables.
3. Apply PCA with varimax rotation [30] to ensure that the principal components rely on the least number of variables (levels).
4. Extract the loadings for the top n principal components, and select m variables (levels) which has the most contributions across all n principal components.

The above procedure shows how dimension reduction works in this study. Note that the number of extracted principal components and the number of levels kept in each categorical predictor are determined by both computational efficiency and prediction accuracy.

Fig.1 shows the entire methodology. It consists of a learning (training) phase and an inference (testing) phase. Together, they form the overall system for predicting the delivery times of current purchase orders (POs) by training our models on historical data for closed POs.

EXPERIMENTAL RESULTS

The experiments were carried out on a MAC High Sierra version 10.13.3 Operating System using R programming language version 3.3.2. A 2.6 GHz Intel Core i7 quad-core processor with 16 GB 2133 MHz LPDDR3 RAM was used. R packages randomForest [31], quantregForest [32] and other supportive packages like ggplot2 [33], reshape2 [34] were employed for

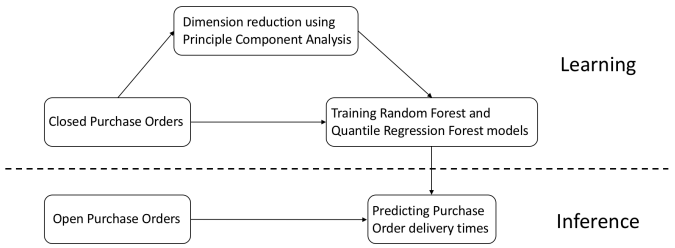

FIGURE 1. FLOW CHART OF PURCHASE ORDERS DELIVERIES PREDICTION METHODOLOGY

implementing Random Forest, Quantile Regression Forest, and results visualization, respectively.

The training dataset consisted of 53,105 closed POs spanning from April, 2016 to May, 2017. The Random Forest and Quantile Regression Forest models were trained using 4 predictors including both discrete and continuous variables. The predictors and their definitions are shown in Tab.1. Further, with the dimension reduction method, 4 more categorical predictors were added to the models. The added predictors and their definitions are shown in Tab.2.

These predictors are selected based on prior knowledge from the OEMs that they are likely to be significant predictors in PO delivery time estimation. For example, different Item IDs have different procurement cycle times, since various materials require different process times, queue times, and procurement times from suppliers. For Quantity, given limited supplier resources, larger order quantities may require longer supplier process times; conversely, in other supplier-OEM arrangements, larger order quantities may be tied to improved prioritization and shorter supplier process times. In case of the ABC Indicator, inventory prioritization at an OEM impacts purchase requisition processes, order priority, and other transactional processes, given that A parts are more critical to an OEM's operations than C parts. For Pdt, the Planned Delivery Time entered into the Enterprise Resource Planning (ERP) system drives several processes, including prompts to Planner-Buyers to create purchase orders, replenishment orders, or VMI orders; it is a critical baseline predictor of the generally acceptable purchase order cycle time for a specific item. Plnt and Vendor are the geographic features of the suppliers, which typically affect the PO cycle times. Material Number and Material Group are also considered to be important variables in the production process.

For dimension reduction, the top 5 principal components were extracted and 15 levels were kept for each categorical variable. All the models were built using 120 trees in the forest, 2 predictors randomly sampled as candidates at each split node, and a minimum of 5 data points in the terminal nodes. 10-fold cross validation was implemented to select the best models. The prediction results are reported in Tab. 3 as the mean absolute errors in days for several percentiles of the POs. For example, the 25th percentile of supplier estimates means that 25% of the POs have prediction errors of less than 2 days using these estimates. The predictions are compared to the supplier estimated delivery dates and a simple linear regression model with the same predictors.

TABLE 1. PREDICTORS USED IN THE RANDOM FOREST AND QUANTILE REGRESSION FOREST MODELS

Predictor	Definition
Item	Item ID
Quantity	Quantity Ordered
ABC Indicator	High-level part classification based on functional requirement
Pdt	Planned delivery time

TABLE 2. PREDICTORS ADDED USING THE DIMENSION REDUCTION METHOD FOR SELECTING THE INFLUENTIAL LEVELS IN THE CATEGORICAL VARIABLES

Predictor	Definition	No. of categories
Plnt	Plant ID	43
Vendor	Vendor ID	881
Material Group	Material Group ID	1,596
Material Number	Material Number ID	4,447

Table 3 shows promising performance of our models in comparison to the supplier estimates and the linear regression model. Compared to the supplier estimates, the linear regression model works better only when the prediction errors are large (90th PO percentile that corresponds to errors of over 27 days), whereas, our models are able to provide more accurate estimates for small prediction errors (50th PO percentile which corresponds to errors of less than 6 days). However, the best estimates for small errors (less than 2 days corresponding to the 25th PO percentile) are provided by the suppliers. This result matches our expectation that for the most familiar or standard POs, likely using regular production capacities, the suppliers provide accurate delivery date estimates. It is, however, very encouraging that even without any direct knowledge of supplier capabilities, our prediction models are superior to their estimates. Quite unsurprisingly, simple linear regression is unable to capture the complex interactions involved in determining the actual PO delivery dates.

Without dimension reduction, Random Forest and Quantile Regression Forest models show comparable performances.

TABLE 3. PREDICTION ERRORS (IN DAYS) AT VARIOUS PO PERCENTILES USING DIFFERENT PREDICTION MODELS AND SUPPLIER ESTIMATES

	25th	50th	75th	90th	95th
Supplier Estimates	**2.00**	6.00	13.00	27.00	41.00
Linear Regression	3.63	7.75	13.26	22.74	35.33
Random Forest w/o Dimension Reduction	2.66	5.91	11.11	20.40	31.25
Quantile Regression Forest w/o Dimension Reduction	2.64	5.88	11.22	20.64	31.60
Random Forest w/ Dimension Reduction	2.28	5.07	**9.77**	**17.98**	**26.95**
Quantile Regression Forest w/ Dimension Reduction	2.27	**5.04**	9.90	18.45	27.95

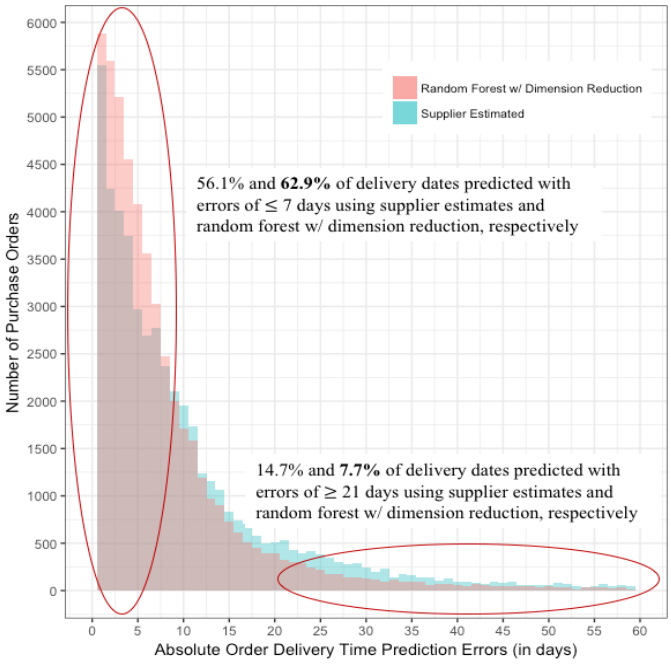

FIGURE 2. HISTOGRAM OF PO DELIVERY DATES PREDICTION ERRORS (LESS THAN 60 DAYS) USING RANDOM FOREST MODEL WITH DIMENSION REDUCTION AND SUPPLIER ESTIMATES

While Quantile Regression Forest is slightly more accurate up to 50th PO percentiles, Random Forest gives better results for higher PO percentiles. A similar trend is seen for the models with dimension reduction. In general, dimension reduction method is effective in providing more accurate predictions as we move toward higher PO percentiles. Our best model improves the prediction accuracy by more than 30% for large lead time POs corresponding to unreliable supplier estimates.

To summarize, Random Forest with dimension reduction has the best overall performance, while Quantile Regression Forest with dimension reduction has marginally better performance for POs with small prediction errors. All our models perform better than supplier estimates and linear regression, except for about a quarter of the POs where the supplier estimates are extremely reliable.

These trends are further reinforced in the prediction errors histogram plots shown in Fig. 2 and Fig. 3. The distributions show that both our models completely encompass the supplier estimates until we reach POs with prediction errors of more than a week. Correspondingly, the supplier estimates show heavy-tailed distributions with substantially more PO delivery dates being predicted inaccurately, including almost 15% with more than 3 weeks of prediction errors.

CONCLUSIONS

In this paper, we address the problem of accurate prediction of POs delivery times in a typical supply chain characterized by a large number of distributed suppliers. We present two regression models in the form of Random Forests and Quantile Regression Forests, which make use of both continuous and categorical predictors, and provide expected means and quantiles of the delivery times, respectively. The models are applied in conjunction with a PCA-based dimension reduction method to select the suitable levels for the categorical predictors.

Testing on a real-world OEM dataset, more than 92%of the POs are shown to have prediction errors of less than three weeks, a commonly used benchmark in OEMs, for lead times of hundreds of days. Furthermore, our regression models outperform the supplier estimates and a simple linear regression fit by 24% for 95% of the POs. After including the dimension reduction method, this performance gain increases to 34%. Therefore, our method provides the OEMs a way to better estimate the POs delivery dates with minimal information and communication from the suppliers. This is a substantial operational benefit for the OEMs, as prediction accuracy is improved by using historical transactional data only, yielding a viable path for optimal inventory management and assembly scheduling.

In the future, we intend to investigate the performance of the

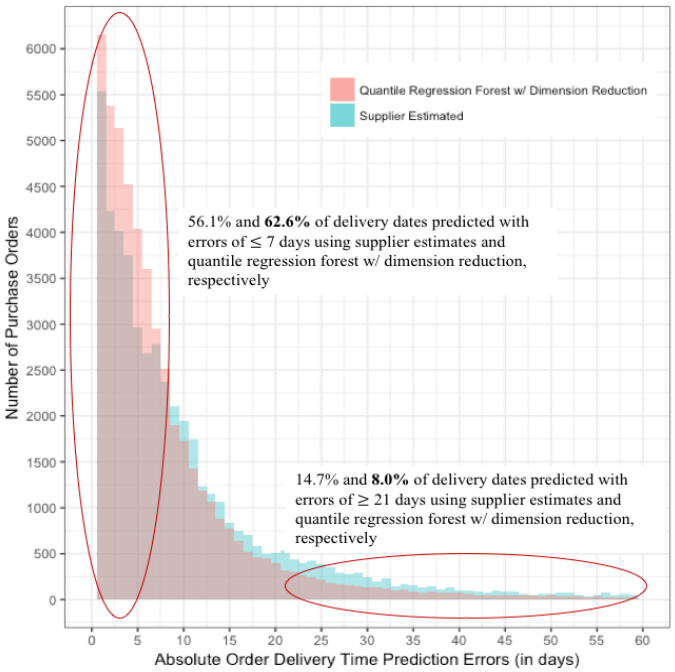

FIGURE 3. HISTOGRAM OF PO DELIVERY DATES PREDICTION ERRORS (LESS THAN 60 DAYS) USING QUANTILE REGRESSION FOREST WITH DIMENSION REDUCTION AND SUPPLIER ESTIMATES

Quantile Regression Forest model using quantiles other than the median as more or less conservative (statistical) estimation may yield better predictions. We then plan to parallelize the process of training the regression models and the dimension reduction method. Such parallel implementation would help us to train the models within a few minutes even for large datasets, and, thereby, enable real-time predictions based on continually updated PO delivery data. It would also be important to evaluate our method on suppliers' datasets, which might contain more detailed production plans and operation sequences as compared to the OEMs' datasets. Considering that both the suppliers and OEMs share the same goal of accurately predicting the completion times of parts orders (either as sales or purchases), we aim to incorporate our models within an easy-to-use visibility tool to facilitate their uses in day-to-day supply chain decision making. We also intend to conduct structured user studies to customize the displays of the models outputs based on the specific needs of the user category.

ACKNOWLEDGMENT

This work is supported by the Digital Manufacturing and Design Innovation Institute (DMDII) through the UI Labs Contract Number 0220160028.

REFERENCES

[1] Banerjee, A. G., Yund, W., Yang, D., Koudal, P., Carbone, J., and Salvo, J., 2015. "A hybrid statistical method for accurate prediction of supplier delivery times of aircraft engine parts". In ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, p. V01BT02A037.

[2] Sra, S., and Dhillon, I. S., 2006. "Generalized nonnegative matrix approximations with bregman divergences". In Advances in Neural Information Processing Systems, pp. 283–290.

[3] Gunasekaran, A., Patel, C., and McGaughey, R. E., 2004. "A framework for supply chain performance measurement". *International Journal of Production Economics,* ***87***(3), pp. 333–347.

[4] Perea-Lopez, E., Ydstie, B. E., and Grossmann, I. E., 2003. "A model predictive control strategy for supply chain optimization". *Computers & Chemical Engineering,* ***27***(8-9), pp. 1201–1218.

[5] Yazdizadeh, P., and Ameri, F., 2015. "A text mining technique for manufacturing supplier classification". In ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, p. V01BT02A036.

[6] Sabbagh, R., and Ameri, F., 2017. "A thesaurus-guided text analytics technique for capability-based classification of manufacturing suppliers". In ASME 2017 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference, American Society of Mechanical Engineers, p. V001T02A075.

[7] Ameri, F., and Dutta, D., 2008. "A matchmaking methodology for supply chain deployment in distributed manufacturing environments". *Journal of Computing and Information Science in Engineering,* ***8***(1), p. 011002.

[8] Zhao, X., Xie, J., and Lau, R., 2001. "Improving the supply chain performance: use of forecasting models versus early order commitments". *International Journal of Production Research,* ***39***(17), pp. 3923–3939.

[9] Zhao, X., Xie, J., and Leung, J., 2002. "The impact of forecasting model selection on the value of information sharing in a supply chain". *European Journal of Operational Research,* ***142***(2), pp. 321–344.

[10] Chen, F., Drezner, Z., Ryan, J. K., and Simchi-Levi, D., 2000. "Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information". *Management Science,* ***46***(3), pp. 436–443.

[11] Chandra, C., and Grabis, J., 2005. "Application of multi-steps forecasting for restraining the bullwhip effect and improving inventory performance under autoregressive demand". *European Journal of Operational Research,* ***166***(2), pp. 337–350.

[12] Aburto, L., and Weber, R., 2007. "Improved supply chain management based on hybrid demand forecasts". *Applied Soft Computing,* ***7***(1), pp. 136–144.

[13] Carbonneau, R., Laframboise, K., and Vahidov, R., 2008. "Application of machine learning techniques for supply chain demand forecasting". *European Journal of Operational Research,* ***184***(3), pp. 1140–1154.

[14] Zamar, D. S., Gopaluni, B., Sokhansanj, S., and Newlands, N. K., 2017. "A quantile-based scenario analysis approach to biomass supply chain optimization under uncertainty". *Computers & Chemical Engineering,* ***97***, pp. 114–123.

[15] Nakano, M., 2009. "Collaborative forecasting and planning in supply chains: The impact on performance in Japanese manufacturers". *International Journal of Physical Distribution & Logistics Management,* ***39***(2), pp. 84–105.

[16] Ramanathan, U., and Gunasekaran, A., 2014. "Supply chain collaboration: Impact of success in long-term partnerships". *International Journal of Production Economics,* ***147***, pp. 252–259.

[17] Cao, M., and Zhang, Q., 2011. "Supply chain collaboration: Impact on collaborative advantage and firm performance". *Journal of Operations Management,* ***29***(3), pp. 163–180.

[18] Ho, T. K., 1995. "Random decision forests". In Proceedings of the Third International Conference on Document Analysis and Recognition, Vol. 1, pp. 278–282.

[19] Meinshausen, N., 2006. "Quantile regression forests". *Journal of Machine Learning Research,* ***7***(Jun), pp. 983–999.

[20] Xuan, S., Liu, G., Li, Z., Zheng, L., Wang, S., and Jiang, C., 2018. "Random forest for credit card fraud detection". In IEEE International Conference on Networking, Sensing and Control, pp. 1–6.

[21] Calderoni, L., Ferrara, M., Franco, A., and Maio, D., 2015. "Indoor localization in a hospital environment using random forest classifiers". *Expert Systems with Applications,* ***42***(1), pp. 125–134.

[22] Altendrof, J., Brende, P., and Lessard, L., 2005. "Fraud detection for online retail using random forests". *Technical Report*.

[23] Efron, B., 1992. "Bootstrap methods: another look at the jackknife". In *Breakthroughs in Statistics*. Springer, pp. 569–593.

[24] Portnoy, S., Koenker, R., et al., 1997. "The Gaussian hare and the Laplacian tortoise: computability of squared-error versus absolute-error estimators". *Statistical Science,* ***12***(4), pp. 279–300.

[25] He, X., Ng, P., and Portnoy, S., 1998. "Bivariate quantile smoothing splines". *Journal of the Royal Statistical Society: Series B (Statistical Methodology),* ***60***(3), pp. 537–550.

[26] Koenker, R., Ng, P., and Portnoy, S., 1994. "Quantile smoothing splines". *Biometrika,* ***81***(4), pp. 673–680.

[27] Donoho, D. L., et al., 2000. "High-dimensional data analysis: The curses and blessings of dimensionality". *AMS Math Challenges Lecture,* ***1***, p. 32.

[28] Pearson, K., 1901. "LIII. on lines and planes of closest fit to systems of points in space". *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science,* ***2***(11), pp. 559–572.

[29] Kolenikov, S., Angeles, G., et al., 2004. "The use of discrete data in PCA: theory, simulations, and applications to socioeconomic indices". *Chapel Hill: Carolina Population Center, University of North Carolina*, pp. 1–59.

[30] Kaiser, H. F., 1958. "The varimax criterion for analytic rotation in factor analysis". *Psychometrika,* ***23***(3), pp. 187–200.

[31] Liaw, A., and Wiener, M., 2002. "Classification and regression by randomforest". *R News,* ***2***(3), pp. 18–22.

[32] Meinshausen, N., 2017. *quantregForest: Quantile Regression Forests*. R package version 1.3-7.

[33] Wickham, H., 2009. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

[34] Wickham, H., 2007. "Reshaping data with the reshape package". *Journal of Statistical Software,* ***21***(12), pp. 1–20.