Proceedings of the ASME 2015 International Design Engineering Technical Conferences & Computers and Information in Engineering Conference
IDETC/CIE 2015
August 2-5, 2015, Boston, Massachusetts, USA

DETC2015-47605

A HYBRID STATISTICAL METHOD FOR ACCURATE PREDICTION OF SUPPLIER DELIVERY TIMES OF AIRCRAFT ENGINE PARTS

Ashis Gopal Banerjee
General Electric Global Research
Niskayuna, NY, USA

Walter Yund
General Electric Global Research
Niskayuna, NY, USA

Dan Yang
General Electric Global Research
Niskayuna, NY, USA

Peter Koudal
General Electric Global Research
Niskayuna, NY, USA

John Carbone
General Electric Global Research
Niskayuna, NY, USA

Joseph Salvo
General Electric Global Research
Niskayuna, NY, USA

ABSTRACT

Aircraft engine assembly operations require thousands of parts provided by several geographically distributed suppliers. A majority of the operation steps are sequential, necessitating the availability of all the parts at appropriate times for these steps to be completed successfully. Thus, being able to accurately predict the availabilities of parts based on supplier deliveries is critical to minimizing the delays in meeting the customer demands. However, such accurate prediction is challenging due to the large lead times of these parts, limited knowledge of supplier capacities and capabilities, macro-economic trends affecting material procurement and transportation times, and unreliable delivery date estimates provided by the suppliers themselves. We address these challenges by developing a statistical method that learns a hybrid stepwise regression – generalized multivariate gamma distribution model from historical transactional data on closed part purchase orders and is able to infer part delivery dates sufficiently before the supplier-promised delivery dates for open purchase orders. The hybrid form of the model makes it robust to data quality and short-term temporal effects as well as biased toward overestimating rather than underestimating the part delivery dates. Test results on real-world purchase orders demonstrate effective performance with low prediction errors and constantly high ratios of true positive to false positive predictions.

KEYWORDS

Supplier performance prediction, stepwise regression, matrix gamma distribution, aircraft engine parts

INTRODUCTION

Aircraft engine assemblies are complex systems consisting of thousands of parts that are provided by a fixed number of pre-qualified suppliers. As shown in Figure 1, the typical flow of parts in the assembly operation consists of sourcing, i.e., the creation of parts purchase orders followed by waiting periods in the supplier internal order queues and actual manufacturing cycle times, storage in inventory warehouses, accumulation in the assembly plants, actual assembly, and extensive failure and performance testing. Thus, supplier delivery of parts is the first, and potentially the most critical, step in the assembly operation process.

The suppliers are usually spread across multiple geographical regions and are often located quite far away from the manufacturing (warehouse, assembly, and testing) sites. Each supplier also has different capabilities and capacities to provide parts of varying types, qualities, and quantities. For example, certain suppliers almost exclusively provide forged or cast parts in bulk, whereas others supply high-quality machined or sheet metal parts in small numbers. This characteristic makes engine assembly operations heavily dependent on sole-sourced parts, and thereby susceptible to the delinquencies of the suppliers. Furthermore, while a few of the assembly operation steps can be performed in parallel, the majority of the steps are sequential, making it necessary for all the parts to be available at appropriate times for these steps to be completed successfully.

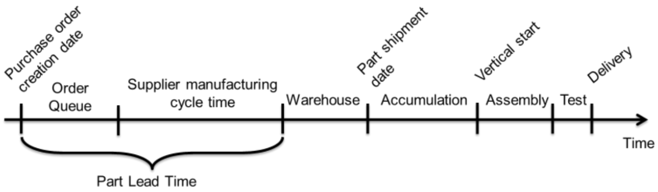


Figure 1. TYPICAL AIRCRAFT ENGINE ASSEMBLY INBOUND MATERIAL (PARTS) FLOW

Accurately predicting the availabilities of the parts enables adequate stocking of inventory parts and ensures uninterrupted assembly operations and modifications of nominal assembly schedules to minimize the delays in meeting the customer demands for the assembled engines. However, such accurate prediction is technically challenging due to the large lead times of these parts of the order of hundreds of days, limited knowledge of time-varying supplier capabilities, and macro-economic trends impacting material procurement and transportation times.

From an operational perspective, the challenge lies in the complete reliance on supplier-provided promise dates, where the promise dates are frequently inaccurate and sometimes unavailable. For example, suppliers may continuously update their promise dates and overwrite the promise date field, eliminating records of their original commitments to deliver the parts. In other cases, suppliers may never update their promise dates, and thus never communicate to the manufacturers (both inventory managers and assembly operators) if there are anticipated delivery problems for a specific part. In general, the promise dates work well as a predictor of supplier delivery dates when the actual delivery dates are close to the requirement dates of the manufacturer. However, if the supplier is delivering parts much earlier or later than originally requested, the promise dates become a poor predictor of when the manufacturer should expect to receive the parts.

We address both the technical and operational challenges by developing a novel method that learns a hybrid stepwise regression – generalized multivariate (matrix) gamma distribution model from historical transactional data on closed part purchase orders. The model is then used to infer part delivery dates several weeks before the supplier-promised delivery dates for open purchase orders to provide sufficient time for remedial actions. The model takes into account three types of predictive factors that are of mixed continuous and categorical types: the dates involved with the purchase orders, surrogate measures of supplier capabilities, and part attributes. The hybrid form of the model enables it to be robust to data quality and short-term temporal effects as well as biased toward conservative forecasts, i.e., it is more likely to overestimate rather than underestimate the part delivery dates.

Test results on real-world purchase orders demonstrate effective performance: prediction errors are low and high ratios of true positive to false positive predictions are constantly achieved. Thus, by applying this hybrid prediction method, the manufacturer can use a small amount of data to adjust the planned engine assembly schedule, and sourcing managers can prioritize corrective actions to ensure that the flow of material to a manufacturing site remains undisrupted.

RELATED WORK

Our work is related to the broad areas of supply chain forecasting, performance measurement, and supplier-customer collaboration, all of which have received a lot of attention over the past several years. While to the best of our knowledge, no specific work has been done on predicting (forecasting) supplier part delivery times using stepwise regression or non-Gaussian probability distributions for large-scale assembly operations, we summarize the vast literature in the abovementioned areas to place our contributions in the proper context.

Unlike our method and application of interest, existing supply chain forecasting models are developed for product demands and inventory stocks, and mostly rely on linear functions of input and output factors with Gaussian noise terms. One of the earliest works on the use of demand forecasting models as a substitute for manufacturer early order commitments to retailers is found in [1]. A subsequent study on the impact of model selection in the context of retailer demand forecasts and inventory replenishment and supplier production decisions is seen in [2]. Autoregressive model based multi-step demand forecasting is applied to improve inventory performance in [3]. A case study of demand forecasting method selection for a global manufacturer of specialty chemical products is done in [4]. Theoretical and empirical analysis of ARIMA demand forecasting model in a manufacturer-retailer supply chain with potential benefits of forecast information sharing is carried out in [5]. Applications of several machine learning methods, including neural networks, recurrent neural networks, and support vector machines on demand forecasting are investigated in [6].

One of the first frameworks for supplier performance measurement from historical data is developed in [7] using supplier rankings based on key performance indicators (KPIs). A review of various performance measures, including non-financial indicators, for logistics operations is provided in [8]. A balanced scorecard method for evaluating supplier performance from four different perspectives, comprising finance, customer, business process, and learning and growth, is presented in [9]. Other representative recent works include essential KPI development [10], common performance measurement system (PMS) for collaboration [11], contextual analysis of PMS [12], and discussion of measurement metrics from strategic, tactical, as well as operational perspectives [13]. In this work, we use a simple performance measure of the accuracy in predicting the parts delivery dates for open purchase orders, which is well suited to our target application.

Of late, there is a lot of interest in collaborative forecasting and planning in supply chains. A study on the impact of such external and internal collaboration is done in [14]. The importance of information quality on the effectiveness of collaboration is investigated in [15]. The impact of collaboration on firm performance in uncertain environments is examined in [16]. In our work, as mandated by real-world operational constraints, there is partial collaboration between the suppliers and manufacturers, where the suppliers provide initial part delivery promise dates but do not share information on their internal capacities.

TECHNICAL APPROACH

Having set the context of our work, we now present the technical details of our prediction system that consists of three steps: a) pre-processing data on supplier part delivery dates over a period of a few years as the input, b) constructing an offline statistical model to predict the delivery dates in a computationally efficient manner, and c) using the model for online predictions on open purchase orders. We first describe the methods for each of these steps and then discuss the overall system architecture. With respect to mathematical notation, scalar variables are represented as regular, lower case English symbols, vector variables as bold, lower case English symbols, matrix variables as bold, upper case English symbols, scalar parameters as regular, lower case Greek symbols, and matrix parameters as regular, upper case Greek symbols.

Data Pre-Processing

Based on the historical transactional data on closed part purchase orders (POs) for a given supplier, we generate a training set of N independent and identically distributed (i.i.d.) paired time series observations (data points) $\boldsymbol{D}_{tr} = (\boldsymbol{X}_{tr}, Y_{tr})$ with independent factor (predictor) variable set $\boldsymbol{X}_{tr} = \{\boldsymbol{x}_1, \dots, \boldsymbol{x}_N\}$, where

$$\boldsymbol{x}_i = ((\boldsymbol{x}_i^o)^T, (\boldsymbol{x}_i^a)^T) = \left(x_{i1}^o, \dots, x_{ip}^o, x_{i1}^a, \dots, x_{ik}^a\right)^T \tag{1}$$

and dependent scalar response variable set $Y_{tr} = \{y_1, \dots, y_N\}$. The factor variables are all positive and real and are of two types: p continuous factors representing the dates associated with the POs and surrogate measures of supplier capabilities such as purchase quantities or costs, and k categorical factors denoting the part attributes such as material types and qualities. The continuous and categorical factors are represented using the superscripts o and a, respectively. The response variable is the delivery date for a particular supplier-part combination.

We first sort the observations so that they are arranged in a non-decreasing order of part PO placement dates. If t_i represents the time instant at which the i^{th} PO is placed in the sorted list, assuming that the 1^{st} PO is placed at time $t_1 = 0$, we weight each observation by a simple normalized coefficient w_i given by

$$w_i = \frac{t_i}{\sum_{i=1}^{N} t_i}\,. \tag{2}$$

This weighting enables us to bias our statistical prediction model toward more recent observations, and better capture any changes in the supplier capabilities. We then scan through the observation set to address data quality issues in the form of negative or missing continuous factors and response variable values. On identifying such issues, the negative values are replaced and the missing values for the factors are filled by the values of the most similar observation in the training set, where the similarity s_{ij} between any two observations $\boldsymbol{d}_i$ and $\boldsymbol{d}_j$, $\boldsymbol{d}_i, \boldsymbol{d}_j \in \boldsymbol{D}_{tr}, i \neq j$, is measured using the positive and non-missing factors and response variable as

$$s_{ij} = \sqrt{\sum_k \left(x_{il} - x_{jl}\right)^2 + \left(y_i - y_j\right)^2}, \tag{3}$$
$$x_{il} \in \boldsymbol{x}_i^o, \qquad x_{jl} \in \boldsymbol{x}_j^o, \qquad x_{il}, x_{jl}, y_i, y_j > 0.$$

Observations are discarded from the training set in the rare instances when a majority of the factors are negative or missing.

The next step is to determine if there is any degree of correlation between the continuous factors and the response variable for a particular part type so as to merit constructing a prediction model for this supplier-part combination. We use the Pearson correlation coefficient for this purpose that is computed for any continuous factor x_j^0 as

$$R_j = \frac{\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}\left(x_{ij} - \bar{\boldsymbol{x}}_j\right)^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}, \tag{4}$$
$$\bar{\boldsymbol{x}}_j = \frac{1}{N}\sum_{i=1}^{n} x_{ij}, \quad x_{ij} \in \boldsymbol{x}_i^o, \quad \bar{y} = \frac{1}{N}\sum_{i=1}^{n} y_i,$$

where $n \leq N$ is the number of observations for the combination being considered. We build the prediction model if R_j is greater than a user-defined threshold for at least one factor x_j^0.

The last step in pre-processing the training data set is to cluster the observations $\boldsymbol{D}_{tr}$ into disjoint groups $\{\boldsymbol{D}_1, \dots, \boldsymbol{D}_S\}$ that span the complete set of observations together so as to construct separate prediction models for each of the groups. Mathematically, this step is formulated as

$$\boldsymbol{D}_{tr} = \cup_{s=1}^{S} \boldsymbol{D}_s\,, \quad \boldsymbol{D}_i \cap \boldsymbol{D}_j = \phi\,,\ \forall i, j\,,\ i \neq j\,. \tag{5}$$

The clusters are generated such that the mean of the response variable values in a cluster is greater or less than the corresponding mean in another cluster by at least a user specified percentage δ. Formally, this condition implies that

$$\frac{|\overline{y_{s+1}} - \overline{y_s}|}{\overline{y_s}} > \delta\,, \qquad \overline{y_s} = \frac{1}{|Y_s|}\prod_{j=1}^{|Y_s|} y_j\,, \tag{6}$$
$$y_j \in Y_s\,, \qquad s = 1, \dots, S-1\,, \qquad \boldsymbol{D}_s = (\boldsymbol{X}_s, Y_s).$$

The test set $\boldsymbol{D}_{te}$, of course, does not contain the response variable. It is simply equal to $\boldsymbol{X}_{te} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M\}$ corresponding to *M* open part POs for which we need to forecast the unknown supplier delivery dates $Y_{te} = \{y_1, \ldots, y_M\}$. We do not apply any temporal weighting to the elements in the test set but insert or replace values for the missing or negative continuous factor elements in exactly the same way as in the training set.

We are now ready to develop the statistical model to predict the supplier delivery dates as a function of the part PO factors. The model is hybrid with two distinct components: a stepwise regression model and a generalized multivariate gamma distribution model, each with its own set of benefits. The regression model is described first followed by the gamma distribution model.

Statistical Modeling: Learning

Stepwise Regression

We use the generalized linear form of the regression function that is given by

$$y_r(\boldsymbol{x}) = b_0 + \sum_{i=1}^{p} b_i^o \boldsymbol{x}_i^o + \sum_{i=1}^{k} b_i^a \boldsymbol{x}_i^a + \sum_{i=1}^{p} c_i^o (\boldsymbol{x}_i^o)^2 + \sum_{i=1}^{p-1} \sum_{j=i+1}^{p} d_{ij}^o \boldsymbol{x}_i^o \boldsymbol{x}_j^o , \quad (7)$$

where $b_0, \{b_i^o\}, \{b_i^a\}, \{c_i^o\},$ and $\{d_{ij}^o\}$ are constant coefficients that are estimated from the training set observations $(\boldsymbol{X}_{tr}, Y_{tr})$ using least squares fitting. Note here that we consider linear, quadratic, and bilinear (interaction) terms for the continuous factors, and just the linear terms for the categorical factors as interaction and higher order effects are usually not observed for the categorical factors given by part attributes. Such a form of the regression model is both compact enough to enable computationally efficient construction and inference, and powerful enough to be able to provide accurate predictions. To further enhance the computational efficiency and predictive power, we adopt the procedure of stepwise elimination of variables [17] to remove the redundant terms from the model that do not contribute significantly in predicting the correct response values.

The stepwise elimination procedure works in the following manner. We begin with the complete model shown above. At each step, we compute the *p*-value of an *F*-statistic on two models, one with all the terms retained so far, and the other with one of the terms removed from the model starting from the interaction terms and progressing to the quadratic and lastly the linear terms. The *F*-statistic is a measure of the importance of a factor (in our case, the extra term present in the 1st model) in explaining the predictive power, i.e., achieving low response variable estimation error, of the model. It is given by

$$F = \frac{\left(\frac{RSS_1 - RSS_2}{m_2 - m_1}\right)}{\left(\frac{RSS_2}{N - m_2}\right)}, \quad RSS_j = \sum_{i=1}^{N} \left(y_i - y_{ij}\right)^2, \ j = 1,2 \quad (8)$$

where m_1 and m_2 are the number of terms in the 1st and 2nd model, respectively, and y_{ij} is the estimated value for y_i using the j^{th} model. If the computed p value is greater than a threshold, typically set as 0.05, then the extra term in the 1st model is removed permanently, and this truncated model becomes the new baseline model at the next step. The removal of this extra term is justified based on the fact that the null hypothesis of having a zero coefficient for this term cannot be rejected. We terminate the procedure when no more terms can be eliminated and consider the trimmed model at the last step as the final regression model.

Matrix Gamma Distribution Fitting

The second component of our hybrid prediction model is a generalized multivariate gamma distribution [18], or matrix gamma distribution in short, that is fitted to the training set observations. The choice of the gamma distribution is based on the fact that it has been shown to be effective for modeling waiting times in econometrics and materials testing as it a more general form of the exponential and chi-squared distributions. Note that the Gaussian distribution would not a suitable model choice for such temporal problems.

Furthermore, the gamma distribution finds widespread use in Bayesian statistics, a branch of statistics that has been very successful in drawing meaningful inferences on the true state of the world (actual part availabilities for aircraft assembly operations in our case) based on belief distributions that are computed using the Bayes' theorem of probabilities. In Bayesian inference, the gamma distribution often serves as the conjugate prior distribution of the unknown parameters of exponential or even other gamma distributions that need to be estimated from a given set of observations [19]. A prior probability distribution is termed as being conjugate to the likelihood function when the posterior (conditional probability after evidence is accounted for) and prior (probability before evidence is taken into account) distributions belong to the same family of distributions.

The probability density function (pdf) of the matrix gamma distribution $f(\boldsymbol{X})$ for a symmetric positive definite matrix $\boldsymbol{X}$ of $\frac{p(p+1)}{2}$ random variables is given by

$$f(\boldsymbol{X}) = \frac{|\Sigma^{-1}|^{\alpha} |\boldsymbol{X}|^{\alpha - \frac{1}{2}(p+1)}}{\beta^{p\alpha} \Gamma_p(\alpha)} \exp\left(-\frac{1}{\beta} tr(\Sigma^{-1}\boldsymbol{X})\right), \ \boldsymbol{X} > 0, \quad \alpha > \frac{(p-1)}{2}, \quad \beta > 0, \quad (9)$$

where $|.|$ denotes the determinant and tr denotes the trace of a matrix, respectively. The multivariate gamma function is defined in terms of the univariate gamma function as

$$\Gamma_{\boldsymbol{p}}(\alpha) = \pi^{\frac{p(p-1)}{4}} \prod_{j=1}^{p} \Gamma\left(\alpha - \frac{1}{2}(j-1)\right) ;$$
$$\Gamma(t) = \int_{0}^{\infty} z^{t-1} \exp(-z) dz \,. \qquad (10)$$

The distribution has three parameters: a shape parameter α, a scale parameter β, and a scale matrix Σ that is symmetric, positive definite, and of full rank. $\boldsymbol{X}$ is then said to follow a matrix gamma distribution denoted as $MG_p(\alpha,\beta,\Sigma)$ and $\boldsymbol{W} = \boldsymbol{X}^{-1}$ follows an inverse matrix gamma distribution $InvMG_p(\alpha,\beta,\Phi)$ with pdf $g(\boldsymbol{W})$ given by

$$g(\boldsymbol{W}) = \frac{|\Phi|^{\alpha}|\boldsymbol{W}|^{-\alpha-\frac{1}{2}(p+1)}}{\beta^{p\alpha}\Gamma_p(\alpha)} \exp\left(-\frac{1}{\beta} tr(\Phi \boldsymbol{W}^{-1})\right), \ \boldsymbol{W},\Phi > 0,$$
$$\alpha \geq \frac{(p-1)}{2} \,, \qquad \beta \geq 0. \qquad (11)$$

In our case, $\boldsymbol{X}$ has the following form

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_{11} & \cdots & \boldsymbol{x}_{1p} \\ \vdots & \ddots & \vdots \\ \boldsymbol{x}_{1p} & \cdots & \boldsymbol{x}_{pp} \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^o & \cdots & \sqrt{\boldsymbol{x}_1^o \boldsymbol{x}_p^o} \\ \vdots & \ddots & \vdots \\ \sqrt{\boldsymbol{x}_1^o \boldsymbol{x}_p^o} & \cdots & \boldsymbol{x}_p^o \end{bmatrix},$$
$$\boldsymbol{x}_i^o > 0 \ \forall \ i \,, \qquad (12)$$

where the matrix elements are either the continuous factors themselves or square root functions of the pairwise products of the continuous factors that account for the interaction effects between the factors.

The next objective is to estimate all the parameters of the matrix gamma distribution based on the training set observations. We adopt a two-pronged approach for this purpose: a maximum likelihood (ML) estimation method [20] for α and β, and a maximum a posteriori (MAP) estimation method [21] for Σ.

We first describe how to estimate α and β, followed by the procedure to estimate Σ. For estimation of α and β parameters, we begin with the univariate gamma distribution that is assumed to model the response variable y. The pdf of this distribution $f(y)$ is then given by

$$f(y) = \frac{y^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)} \exp\left(-\frac{y}{\beta}\right), \ \ y > 0, \qquad (13)$$

and y is said to follow a Gamma distribution denoted by $G(\alpha,\beta)$. Note that we assume the same shape and scale parameters for the matrix and the univariate Gamma distributions as the factor variables and the response variable are intrinsically paired together in the same observation set. Now, the likelihood function for $f(y)$ is given by

$$\begin{aligned} &L(\alpha,\beta) \\ &= \prod_{i=1}^{N} f(y_i) \\ &= \frac{1}{\beta^{N\alpha}|\Gamma(\alpha)|^N} \left(\prod_{i=1}^{N} y_i^{\alpha-1}\right) \exp\left(-\frac{1}{\beta}\sum_{i=1}^{N} y_i\right), \end{aligned} \qquad (14)$$

from which we compute the log-likelihood function as

$$\begin{aligned} l(\alpha,\beta) &= \ln\big(L(\alpha,\beta)\big) \qquad (15) \\ &= (\alpha-1)\sum_{i=1}^{N} \ln(y_i) - \beta \sum_{i=1}^{N} y_i + N\alpha \ln(\beta) - N \ln\big(\Gamma(\alpha)\big). \end{aligned}$$

Taking the partial derivatives of $l(\alpha,\beta)$ with respect to α and β, and setting them to 0, after some algebraic manipulations, we obtain the ML estimates of α and β, denoted by $\hat{\alpha}$ and $\hat{\beta}$, respectively, as

$$\hat{\beta} = \frac{\hat{\alpha}}{\bar{y}} \text{ and}$$
$$\ln(\hat{\alpha}) - \psi(\hat{\alpha}) = \ln\left(\frac{\bar{y}}{\tilde{y}}\right),$$
$$\psi(\hat{\alpha}) = \frac{\Gamma'(\hat{\alpha})}{\Gamma(\hat{\alpha})}, \quad \tilde{y} = \left(\prod_{i=1}^{N} y_i\right)^{\frac{1}{N}}, \qquad (16)$$

where $\psi(\hat{\alpha})$ is referred to as the digamma function [22]. As the last equation does not permit a closed-form solution for $\hat{\alpha}$, we compute an initial approximate value of $\hat{\alpha}$ using the following expression

$$\hat{\alpha} \approx \frac{3 - r + \sqrt{(r-3)^2 + 24r}}{12r} \,, \quad r = \ln\left(\frac{\bar{y}}{\tilde{y}}\right), \qquad (17)$$

and then update $\hat{\alpha}$ using an explicit form of the Newton-Raphson iteration method until convergence is reached. The iteration step takes the form of

$$\hat{\alpha} \leftarrow \hat{\alpha} - \frac{\ln(\hat{\alpha}) - \psi(\hat{\alpha}) - r}{\frac{1}{\hat{\alpha}} - \psi'(\hat{\alpha})}. \qquad (18)$$

We now use the estimates of α and β to derive a MAP estimate for the scale matrix Σ using an analytical method that considers Σ as an observation statistic-dependent matrix of random variables. Selecting the matrix of training set factor means $(\overline{\boldsymbol{X}_{tr}})$ as the observation statistic, the sampling distribution or likelihood function $f(\overline{\boldsymbol{X}_{tr}}|\Sigma)$ of $\overline{\boldsymbol{X}_{tr}}$ is assumed to follow a matrix gamma distribution $MG_p\big(\hat{\alpha},\hat{\beta},\Sigma\big)$, where

$$\overline{\boldsymbol{X}_{tr}} = \begin{bmatrix} \overline{\boldsymbol{x}_1^o} & \cdots & \overline{\sqrt{\boldsymbol{x}_1^o \boldsymbol{x}_p^o}} \\ \vdots & \ddots & \vdots \\ \overline{\sqrt{\boldsymbol{x}_1^o \boldsymbol{x}_p^o}} & \cdots & \overline{\boldsymbol{x}_p^o} \end{bmatrix}, \qquad \overline{\boldsymbol{x}_j^o} = \frac{1}{N} \sum_{i=1}^{N} x_{ij}^o \, ,$$

$$\overline{\sqrt{\boldsymbol{x}_l^o \boldsymbol{x}_m^o}} = \frac{1}{N} \sum_{i=1}^{N} \sqrt{x_{il}^o x_{im}^o} \, . \tag{19}$$

When the likelihood distribution is matrix gamma with known shape and scale parameters but unknown scale matrix, then both the posterior and conjugate prior distributions of Σ follow the inverse matrix gamma distribution [21]. Supposing that the prior distribution $g(\Sigma)$ is given by $InvMG_p(a, \hat{\beta}, \Phi)$, $a \geq \frac{p-1}{2}$, $\hat{\beta} \geq 0$, the posterior distribution $f(\Sigma | \overline{\boldsymbol{X}_{tr}})$ then follows

$$InvMG_p\left(\hat{\alpha} + a, \hat{\beta}, \overline{\boldsymbol{X}_{tr}} + \Phi\right),$$
$$\hat{\alpha} + a \geq \frac{p-1}{2}, \qquad \hat{\beta} \geq 0, \qquad (\overline{\boldsymbol{X}_{tr}} + \Phi) > 0. \tag{20}$$

.

The MAP estimator of Σ, $\hat{\Sigma}_{MAP}$, is then given by the mode of its posterior distribution [21], which is also equal to the Bayes estimator using a cross-entropy based loss function [23] of the form

$$L_{CE}\left(\Sigma, \hat{\Sigma}_{MAP}\right) = \ln\left(\frac{f(\Sigma | \overline{\boldsymbol{X}_{tr}})}{f\left(\hat{\Sigma}_{MAP} | \overline{\boldsymbol{X}_{tr}}\right)}\right). \tag{21}$$

Using the Bayes' theorem, $\hat{\Sigma}_{MAP}$ is found to be

$$\hat{\Sigma}_{MAP}(\overline{\boldsymbol{X}_{tr}}) = \arg\max_{\Sigma} f(\Sigma | \overline{\boldsymbol{X}_{tr}}) = \arg\max_{\Sigma} f(\overline{\boldsymbol{X}_{tr}} | \Sigma) g(\Sigma)$$
$$= \frac{(\overline{\boldsymbol{X}_{tr}} + \Phi)^{-1}}{\hat{\beta}\left\{(\hat{\alpha} + a) + \frac{1}{2}(p+1)\right\}}. \tag{22}$$

We choose Φ to be equal to the identity matrix of order p, $I_{p \times p}$, and $a = \frac{p-1}{2}$ to complete the parameter estimation procedure.

Statistical Modeling: Inference

The last component of our hybrid prediction system is the actual inference method that uses the stepwise regression and matrix gamma distribution models to predict the unknown supplier part delivery dates y_{te} for the test set of observations $\boldsymbol{X}_{te}$. While it is straightforward to use the regression model for prediction by simply inserting the factor variable observations $\boldsymbol{x}_i \in \boldsymbol{X}_{te}$ as $\boldsymbol{x}$ in $y_r(\boldsymbol{x})$, the process is non-intuitive for the matrix gamma model that provides a pdf $f(\boldsymbol{X})$ instead of a directly computable form of y_{te}.

To generate the predictions using the matrix gamma model, we construct a $\left(\frac{p(p+1)}{2} + 1\right)$ dimensional look-up table of probabilities $P_g(\boldsymbol{X}, y)$ for known factor and response variable values in the training set $(\boldsymbol{X}_{tr}, Y_{tr})$ from $MG_p\left(\hat{\alpha}, \hat{\beta}, \hat{\Sigma}_{MAP}\right)$. For a given $\boldsymbol{X}_{te}$, the predicted value of y, denoted as $y_g(\boldsymbol{X}_{te})$, is then obtained by searching over the space of y so as to minimize the absolute difference between the probability returned by the pdf $f(\boldsymbol{X}_{te})$ and the linearly interpolated value from the look-up probability table. This step is mathematically represented as

$$y_g(\boldsymbol{X}_{te}) = \arg\min_y \left| f(\boldsymbol{X}_{te}) - P_g(\boldsymbol{X}_{te}, y) \right|. \tag{23}$$

The final step in prediction is to compute y_{te} by taking the maximum of the values returned by the individual models so as to obtain conservative predictions that are more likely to estimate late rather than early part delivery dates. Mathematically, this step amounts to

$$y_{te} = \max\left(y_r(\boldsymbol{X}_{te}), y_g(\boldsymbol{X}_{te})\right). \tag{24}$$

Prediction System

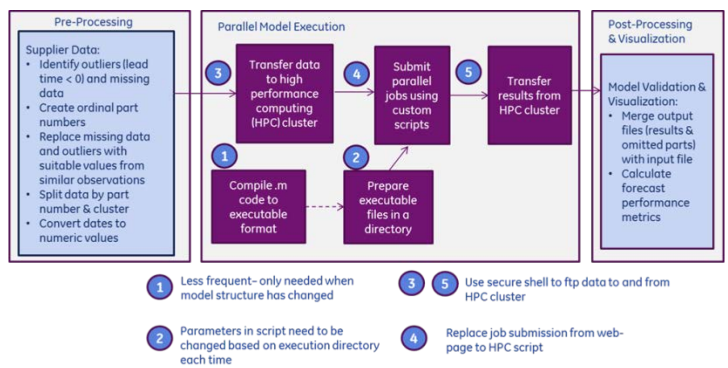


Figure 2. SUPPLIER PART DELIVERY DATE PREDICTION SYSTEM DIAGRAM

Figure 2 illustrates all the components and methods used in these components in our prediction system. While the methods have already been described in details, it is worthwhile to point out two other salient features of the system. First, the prediction models are developed not just for one but every supplier involved with a particular large-scale assembly operation of interest, and high performance computing (HPC) resources are utilized for constructing all the models in parallel to drastically reduce the offline computation time. Second, the system is almost entirely automated with minimal manual interventions required at a few parallel model execution steps to avoid human errors and time lags between the steps as far as possible.

EXPERIMENTAL RESULTS

We now report the experimental results on the effectiveness of the hybrid prediction system on real-world supplier data associated with aircraft engine assembly operations.

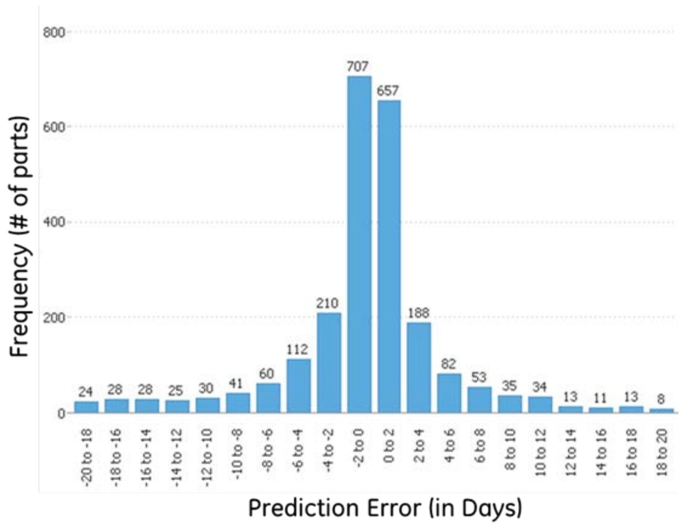
Part-Centric Prediction

As a part of an initial validation exercise, we first present the results using a simpler model, referred to as the part-centric prediction model, which only considers the continuous factors. This simpler model is constructed for each and every part instead on a supplier-by-supplier basis. The training dataset comprised 32 U.S.-based suppliers providing 2,507 parts to assemble a GE Aviation jet engine; about 3,000 clusters were obtained during pre-processing. The models were implemented in MATLAB® R2013a. A separate model was learned for every cluster and the entire learning process took about 14 hours of CPU time on a 1000 core Unix HPC cluster. The actual model construction time was only 3 hours and 49 minutes, with the remaining time spent on data I/O operations and queue waiting periods for the shared HPC resources. The minimum, average and maximum time required to construct a model was 6.33 seconds, 10.56 seconds, and 160.38 seconds, respectively. These values highlight the computational efficiency of our system as compared to simulation or sampling-based methods that often take several minutes to hours for a single model generation.

Table 1. PART-CENTRIC PREDICTION MODEL PERFORMANCE ON A TRAINING DATASET OF CLOSED PURCHASE ORDERS

Model Performance Measure (in days)	Value (error)	Value (absolute error)
Median prediction error	0.05	10.85
Prediction error inter-quartile range	21.70	23.30
Prediction error 10^{th} percentile	-33.34	1.16
Prediction error 90^{th} percentile	33.26	61.28

Table 1 and Figure 3 show promising performance of the system with respect to modeling the parts delivery dates in the training dataset. While a majority of the parts delivery dates are modeled very accurately, a few of the parts show large discrepancies between the modeled and actual delivery dates, leading to high (low) values of the modeling error 90^{th} (10^{th}) percentiles or long tails of the error histogram distribution. We hypothesize that such errors can be significantly reduced by incorporating the part attribute factors in the prediction model. The following results of the complete model on multiple test datasets, comprising of open POs that are distinct from the corresponding training datasets containing closed POs, support this hypothesis and clearly establish the effectiveness of the model.



78% of predicted values are within ±6 days of actual parts delivery dates

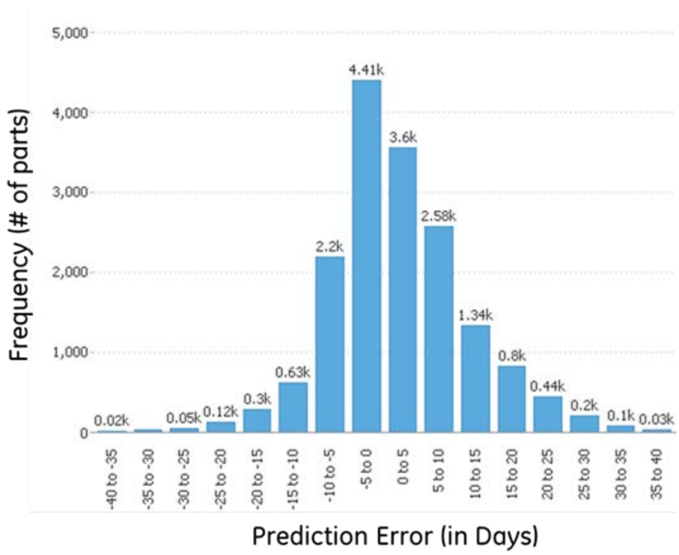
Figure 3: HISTOGRAM OF PART-CENTRIC MODEL PREDICTION ERRORS (DIFFERENCES BETWEEN ACTUAL AND MODELED PARTS DELIVERY DATES) ON A TRAINING DATASET

Supplier-Centric Prediction

Table 2. SUPPLIER-CENTRIC PREDICTION MODEL PERFORMANCE ON TWO LARGE TEST DATASETS OF OPEN PURCHASE ORDERS.

Model Performance Measure (absolute value in days)	Dataset 1	Dataset 2
Median prediction error	6.78	5.39
Prediction error standard deviation	14.43	7.02
Prediction error inter-quartile range	10.05	7.38
Prediction error 10^{th} percentile	1.22	0.95
Prediction error 90^{th} percentile	27.27	16.92

Table 2 and Figure 4 demonstrate the predictive power of our complete model that includes both the continuous and categorical factors, referred to as the supplier-centric model to distinguish it from the simpler part-centric model. The test datasets comprise more than 5,000 parts and the evaluation is done for all the orders that are due at least 6 weeks later. The median prediction errors are small for both the test datasets with values of less than a week for parts having lead times of 200-700 days. Even more importantly, the variabilities (or dispersion measures) in the prediction errors are acceptable and much lower than the corresponding values for the part-centric model. The tight nature of the prediction error distribution is particularly evident in Figure 4 shown for Dataset 2. Note here that the training set used for Dataset 2 is considerably larger than the Dataset 1 (589,130 records vs. 534,764 records) and includes a higher proportion of POs with short lead times. Most of the new additions to Dataset 2 also consist of POs that are delivered right after a fiscal quarter closure, which is typically a period in time when POs that are due to be delivered soon are pushed out to reduce inventory. Due to this reason, the ratio of POs that are delivered on-time relative to their requirements dates in Dataset 2 is markedly higher than Dataset 1, leading to better prediction results.



75% of predictions are within ± 10 days of actual parts delivery dates, 94% are within ± 20 days

Figure 4. HISTOGRAM OF SUPPLIER-CENTRIC MODEL PREDICTION ERROR ON A LARGE TEST DATASET

Table 3. PERFORMANCE EVALUATION OF VARIOUS PREDICTION MODELS ON A LARGE TEST DATASET.

Performance Measure (absolute value in days)	Linear Regression	ARIMA	Hybrid model
Median prediction error	10.35	7.47	**5.39**
Prediction error standard deviation	16.89	13.11	**7.02**
Prediction error inter-quartile range	17.46	12.98	**7.38**
Prediction error 10^{th} percentile	1.53	1.14	**0.95**
Prediction error 90^{th} percentile	35.92	29.25	**16.92**

Table 3 shows that our hybrid prediction model outperforms two other commonly used time series prediction models. The linear regression model only considers the constant bias and linear terms in Eq. (7). The ARIMA model, an acronym for autoregressive integrated moving average model, is a combination of two models, one that accounts for non-stationarity of the data, and the other that captures both the autoregressive nature (the response variable depends linearly on its own prior values) and unobserved white noise errors in the data (moving average) [24]. This result is, thus, particularly encouraging as our model provides higher prediction accuracy using measures of both central tendency and dispersion by avoiding any normality and linearity assumptions on the supplier part delivery dates.

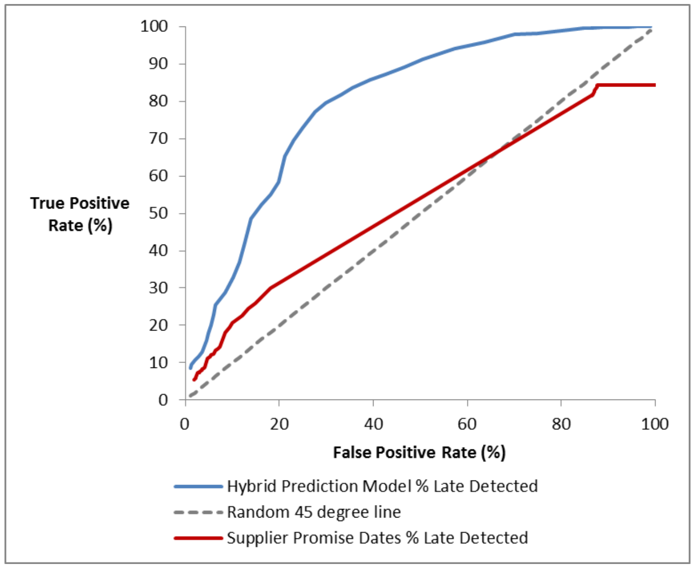


Figure 5. ROC CURVES OF SUPPLIER PARTS DELIVERY DATES PREDICTION MODELS

The final demonstration of the effectiveness of our hybrid prediction model is illustrated using the ROC curves in Figure 5. We define a true positive prediction as one where the model predicts a late part delivery for a part that is actually delivered late by more than a user-specified threshold from the required delivery date. Thus, we obtain various points on the curve by setting this threshold to different values ranging from 1 day to 20 days. Consistently high ratios of the true positive to false positive prediction rates show both high sensitivity and specificity of the model, lending further credence to our claim of the model's operational value. In contrast, supplier provided promise dates perform comparably to that of random predictions highlighting the need for a much more accurate model.

CONCLUSIONS

In this paper, we address the problem of accurate prediction of supplier part delivery dates for aircraft engine assemblies. We develop a hybrid method by combining stepwise generalized linear regression with matrix gamma distribution fitting model for this purpose that does not assume normality of the part delivery distributions or a linear combination of heuristically-chosen predictor factors. Our model automatically selects only the significant factors and interactions among the factors and leverages historical information on supplier performance to estimate the parameters of the matrix gamma distribution.

Testing on real-world parts purchase orders yield delivery time prediction accuracy of the order of a few (5-7) days on average with acceptable standard deviation (< 15 days) for parts having lead times of several hundred days. The ratio of true positive to false positive predictions is constantly high, where positive predictions correspond to the estimated part delivery dates that are late by more than a user-specified threshold from the requirement dates. The method is also computationally efficient due to the estimation of the gamma distribution parameters using analytical or semi-analytical expressions instead of Markov Chain Monte Carlo or other sampling methods, and parallel model learning for all the suppliers or part clusters using HPC resources. Thus, our method provides substantial operational benefits by identifying late arriving parts several weeks in advance of their scheduled delivery dates leading to possible remedial actions in the forms of supplier engagement or stocking of parts from other sources.

In future, we plan to use the model to identify serially delinquent suppliers leading to changes in their qualification status or imposition of more stringent contract terms. We then intend to carry out optimal rescheduling of engine assembly operations based on the predicted part delivery dates and current inventory levels. While weighting the observations helps in modeling the changes in the supplier capabilities over time, we also aim to investigate suitable selection of thresholds to completely filter out old observations that provide inaccurate capability estimates in the prediction model.

ACKNOWLEDGMENTS

We gratefully acknowledge the help and support of Adam Bangert, Mark Grabowski, and Yogesh Srivastava at GE Aviation and Stephan Biller at GE Global Research.

REFERENCES

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

[2] 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***, pp. 321-344.

[3] 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), October, pp. 337-350.

[4] 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), February, pp. 1140-1154.

[5] Acara, Y., and Gardner Jr, E. S., 2012. “Forecasting method selection in a global supply chain”. *International Journal of Forecasting*, ***28***(4), October-December, pp. 842-848.

[6] Babaia, M. Z., Alic, M. M., Boylanc, J. E., and Syntetos, A. A., 2013. “Forecasting and inventory performance in a two-stage supply chain with ARIMA(0,1,1) demand: Theory and empirical analysis”, *International Journal of Production Economics*, 143(2), June, pp. 463-471.

[7] Nagar, A., 2006. *System and method for evaluating supplier performance in a supply chain*. U.S. Patent No. 20060053063 A1.

[8] Gunasekarana, A., and Kobu, B., 2007. “Performance measures and metrics in logistics and supply chain management: a review of recent literature (1995–2004) for research and applications”. *International Journal of Production Research*, ***45***(12), May.

[9] Bhagwata, R., and Sharma, M. K., 2007. “Performance measurement of supply chain management: A balanced scorecard approach”. *Computers & Industrial Engineering*, ***53***(1), August, pp. 43-62.

[10] Chae, B., 2009. "Developing key performance indicators for supply chain: an industry perspective". *Supply Chain Management: An International Journal*, ***14***(6), pp. 422 – 428.

[11] Papakiriakopoulos, D., and Pramatari, K., 2010. "Collaborative performance measurement in supply chain". *Industrial Management & Data Systems*, ***110***(9), pp. 1297 – 1318.

[12] Cuthbertson, R., and Piotrowicz, W., 2011. "Performance measurement systems in supply chains: A framework for contextual analysis", *International Journal of Productivity and Performance Management*, ***60***(6), pp. 583 – 602.

[13] Cho, D. W., Lee, Y. H., Ahn, S. H., and Hwang, M. K., 2012. “A framework for measuring the performance of service supply chain management”. *Computers & Industrial Engineering*, ***62***(3), April, pp. 801-818.

[14] 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.

[15] Wiengarten, F., Humphreys, P., Cao, G., Fynes, B., and McKittrick, A., 2010. "Collaborative supply chain practices and performance: exploring the key role of information quality". *Supply Chain Management: An International Journal*, ***15***(6), pp. 463 – 473.

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

[17] Draper, N. R., and Smith, H., 1998. *Applied Regression Analysis*. Wiley-Interscience, Hoboken, NJ.

[18] Anderson, T., 1984. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, 2nd Edition, Hoboken, NJ.

[19] Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B., 2003. *Bayesian Data Analysis*, CRC Press, 2nd edition, Boca Raton, FL.

[20] Moala, F. A., Ramos, P. L., and Achcar, J. A., 2013. "Bayesian Inference for Two-Parameter Gamma Distribution Assuming Different Noninformative Priors". *Revista Colombiana de Estadística*, ***36***(2), December, pp. 321-338.

[21] Das, S., and Dey. D. K., 2010. "On Bayesian Inference for Generalized Multivariate Gamma Distribution". *Statistics & Probability Letters*, ***80***(19-20), October, pp. 1492-1499.

[22] Lawless, J., 1982. *Statistical Models and Methods for Lifetime Data*. John Wiley, New York, NY.

[23] Murphy, K. P., 2012. *Machine Learning: A Probabilistic Perspective*. The MIT Press, Cambridge, MA.

[24] Dimitros, A., and Hall, S. G., 2011. *ARIMA Models and the Box–Jenkins Methodology*. 2nd Edition of *Applied Econometrics*. Palgrave MacMillan, New York, NY, pp. 265–286.