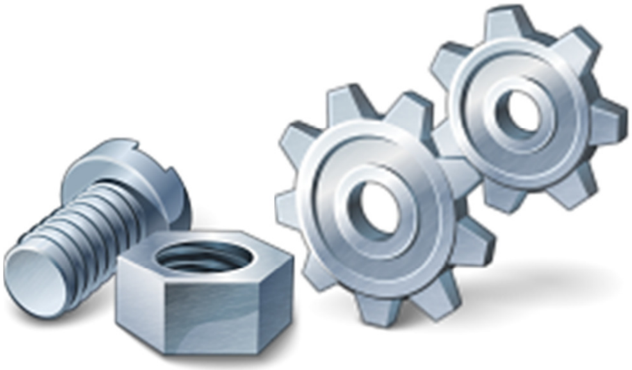
WHITEPAPER

Spare Parts Inventory Management with Quantile Technology



Spare Parts Inventory Management with Quantile Technology

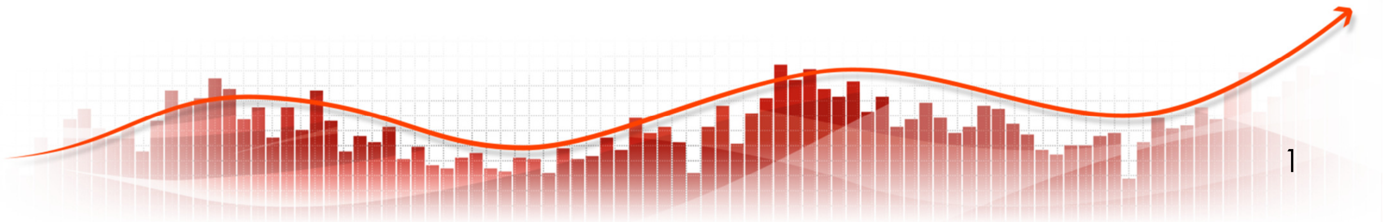
Introduction

The management of spare and service parts is as strategically important as it is difficult. In a world where most equipment manufacturers and retailers are operating in fiercely competitive markets, a high service level to the existing customer base is a strategic priority for many players.

Not only does a high spare part availability help build a loyal base of customers that return to buy new products and acts as a multiplicators in the market. Product/equipment companies have also discovered services as an often very profitable and recurring revenue stream that is typically more resilient to economic cycles than equipment sales.

Siemens' division for power generation once regarded services as a distraction from the business of building equipment. Today, services account for more than 30% of division revenues.

However, managing a spare parts inventory efficiently still poses a huge challenge due to size, service level requirements and nature of demand. This whitepaper discusses the challenges and current state of spare parts planning technology, and introduces quantile forecasting as a disruptive new approach to tackling the problem.



Forecasting and planning spare parts is particularly challenging

Despite a forecasting and inventory planning technology industry that is several decades old, spare parts management has remained a challenge for a number of reasons:

- **Large number of parts**: Even smaller equipment manufacturers can easily be confronted with managing more than a hundred thousand spare parts. Larger OEMs often reach a million spare parts or more. The sheer number of parts to manage makes 'smart' human intervention unrealistic for most companies.
- **High service level requirement:** Stock outs are often very costly as machine downtimes can halt whole production lines and accumulate vast losses of productivity in a short time. High to very high service levels are therefore paramount in many industries, independently who is carrying the risk of a downtime (either the client or the OEM via service level agreements).
- **Infrequent demand:** The demand for spare parts is typically sparse and intermittent, meaning that only very low volumes are required occasionally. Existing 'classic' forecasting and inventory planning methods work very poorly with this type of demand pattern.

Unfortunately, the combination of these factors **makes standard inventory and forecasting technology ill-suited for spare parts planning**.

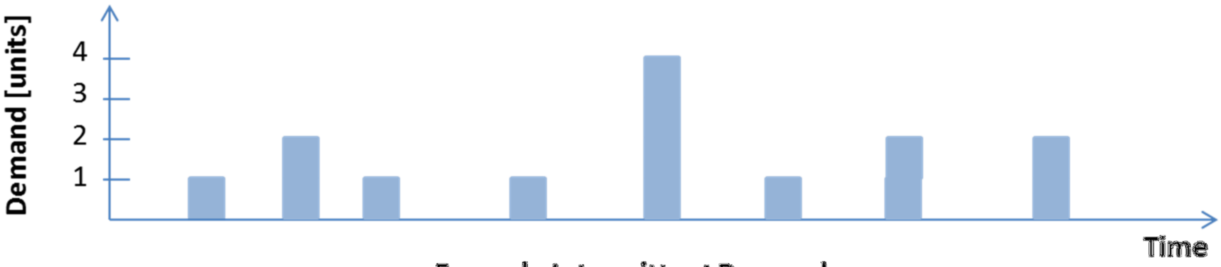
Why standard forecasting technology performs poorly

While individual vendors will differentiate in user interfaces, reporting capabilities and particular features, the industry has been working for decades with the same underlying statistical theory.

"Classic forecasting methods have been developed for high rotation products with 'thick' demand patterns - here they work well. Forecasting slow movers however is a very different problem"

In classic forecasting and inventory planning theory, a forecast is produced by applying models such as moving average, linear regression, Holt Winters or adaptive smooting. A great deal of attention is given to the forecasting error, which is optimized by measuring MAPE or similar indicators. The transformation into a suggested stock level is done in a second step via classic safety stock analysis.

This methodology has been developed decades ago for the application in manufacturing and retail, which are characterized by ‘thick’ time series, i.e. the planning of products that are produced or sold in high numbers and with little intermissions. Here, this method works well.



Example Intermittent Demand

Unfortunately, in the case of sparse time series (also called slow movers: low unit and infrequent sales), this methodology fails. The main issue with forecasting slow movers is that **what we are essentially forecasting are zeros**. This is intuitively obvious when looking at the demand history of a typical spare parts portfolio on a daily, weekly, or even monthly basis: By far the most frequent data point is zero, which can in some cases make up more than 50% of all recorded data points.

The challenge of forecasting slow movers: Good statistical performance and good inventory practice are not the same

When applying classic forecasting theory to this type of data set, the best forecast for a slow moving product is by definition a zero. A 'good' forecast from a statistical point of view will return mostly zeros, which is optimal in terms of math, but not useful in terms of inventory optimization.

The classic method completely separates the forecast from replenishment. The problem is, the situation can hardly be improved with a “better” forecast. All efforts to increase the forecasting accuracy according to a chosen

“When applying classic forecasting theory to slow movers, the best forecast from a statistical perspective is by definition a zero. What is ‘best’ from a theoretical perspective is not useful for inventory optimization”

accuracy criteria are a rather theoretical exercise and unfortunately a battle in the wrong place. **What actually matters in practice is the accuracy of the resulting inventory level, which is not measured nor optimized**.

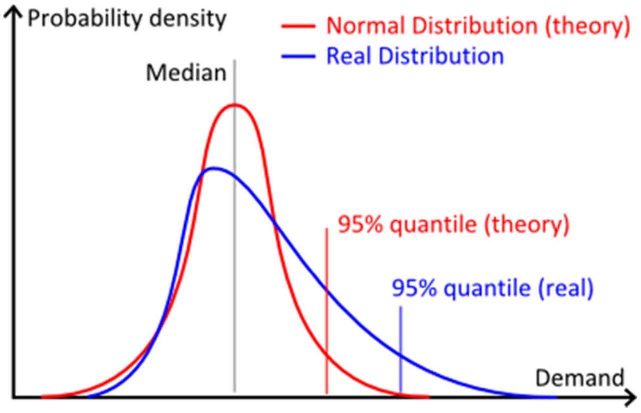
What's worse: Classic (safety stock) theory fails to transform the forecast correctly into inventory policy

An additional problem in classic inventory management theory is the transformation of the forecast into inventory policy, in most cases is the reorder point (minimum inventory level that should trigger a replenishment order) or target stock level.

A classic forecast is by definition a 'mean' forecast, which means the probability of being too high or too low is equal. In other words, the forecast will give a 50% service level, or "50% quantile". By adding safety stock, the service level is then increased to the level which reflects the economic reality of doing business. This is nothing else but 'extrapolating' to a higher quantile (from 50% to e.g. > 90%).

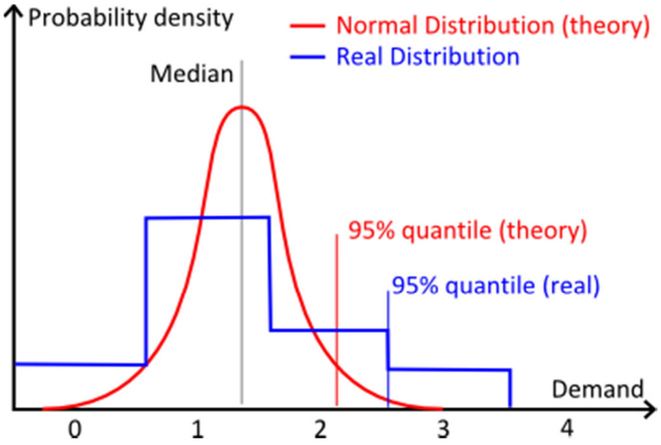
Classic safety stock analysis makes the assumption that demand and error distribution following a normal curve. While this is a good approximation in the case of 'thick' time series and service levels close to the mean, it is a very poor approximation in the case of intermittent demand and high service levels.

High service levels (high quantiles)



The assumption that errors associated to forecasts are normally distributed is typically good for service level targets close to the mean or the median. However, the quality of the approximation degrades as the target service level increases. For high target percentages, typically all values above 90%, the extrapolation itself frequently degrades in quality significantly due to a poor approximation of the reality.

Intermittent demand



In the case of intermittent demand, the extrapolation tries to fit a smooth (normal) curve over the future demand in order to reflect uncertainty. However, when the demand is intermittent or sparse, there is nothing smooth about the demand: for each period (week, month), the number of units being sold, i.e. the observable demand, is an integer varying between 0 and 5 for example.

Historically, many classic (mean) forecasting models have been designed to better work with sparse demand; however it becomes clear that the more fundamental issue is that no mean forecast can be properly extrapolated into an accurate quantile in case of sparse demand.

Changing the vision from Forecast Accuracy to Risk Management

When dealing with slow movers, we believe the right approach is not to approach the problem as a forecasting issue and to try to forecast demand (which is mostly zero).

Much rather, **the analysis should provide an answer to the question how much inventory is needed in order to insure the desired service level.** The whole point of the analysis is not a more accurate demand forecasts, but a better risk analysis. We fundamentally change the vision here.

"Instead of approaching the problem as a forecasting issue, we determine directly how much inventory is needed in order to insure the desired service level."

Determining and optimizing directly the Reorder Point

Quantile forecasts allow the forecasting of the optimal inventory that provides the desired inventory level directly: A bias is introduced on purpose from the start in order to alter the odds of over and under forecasting.

Definition: A quantile forecast (τ, λ) is the minimum inventory level for the lead time λ that provides the service level τ (i.e. a probability τ of being higher than the future demand).

A quantile forecast will therefore take as input the demand (or sales) history, the lead time and the desired service level and directly output the adequate reorder point, thus by-passing the classic forecast optimization plus safety stock analysis.

> **„The quantile forecast accuracy is determined by assessing the accuracy of the reorder point directly, instead of optimizing a demand forecast that is far removed from the reality of inventory requirements"**

The terminology quantile forecast might sound complicated and is little known among supply chain experts. However, quantile forecasts – without being named that way - are routinely used in retail and manufacturing businesses. For example, defining a reorder point for your inventory is strictly equivalent to producing a quantile forecast over the demand.

Calculating the safety stock required to provide a desired service level is nothing else but producing (or rather extrapolating) a quantile forecast, only in a very 'lossy' way.

The accuracy of a quantile forecast is determined by assessing the accuracy of the reorder point directly. Instead of optimizing a forecast that is still far removed from the required inventory policy by using a theoretical accuracy criteria, what matters is assessed directly: **The actual inventory level versus demand.**

Definition: *Extrapolated quantiles* are classic (mean) forecasts transformed into quantile forecasts through an extrapolation (typically safety stock analysis). *Native quantiles* directly produces the quantile via the statistical model.

In order to account for the a-symmetrical cost of being too high or too low (a stock out being much costlier than a unit too many in stock), this is done effectively by using the “pinball loss function” which assigns different ‘scores’ for data points being too low or too high.

It is important to note that quantile forecasts address seasonality, product life cycle, trends and patterns in the same way classic forecasts do.

Performance increase by up to 50% allows reduced inventory and/or higher accuracy

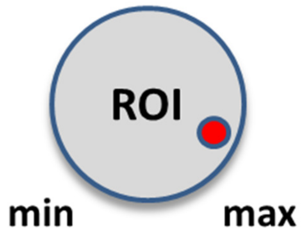
In practice, native quantile forecasting technology will outperform classic theory in most applications, but particularly in the case of high service levels and intermittent demand. It is therefore a technology solution that will particularly improve the performance of spare parts inventory management systems.

Benchmarks against classic forecasting technology in food, non-food, hardware, luxury and spare parts consistently show that quantile forecasts bring a performance improvement of over 25%, that is either more than 25% less inventory or 25% less stock outs.

> **„Performance increase of up to 50% can be quantified in benchmarks. These can translate to significantly reduced inventory and/or a higher availability”**

In the case of spare parts management we typically see a desire of clients to improve service levels while not increasing stock drastically. In a typical application case the performance improvement is significant enough to achieve a higher service level while lowering inventory. The distribution of this performance towards inventory level and availability can be finely tuned via the choice of target service level.

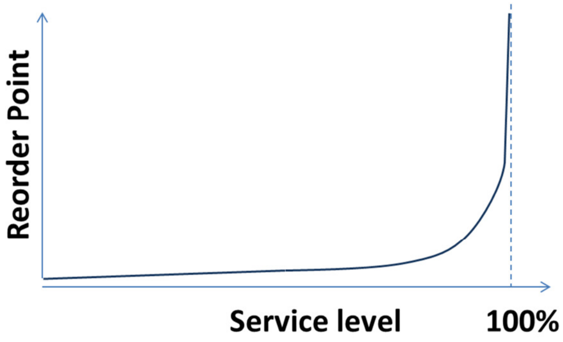
Differentiated service levels unlock further potential



Further potential lies in the fact that not all parts are equally critical to customers, and that the amount of stock required to deliver the same service level will vary strongly from part to part.

The higher the demand volatility, and therefore uncertainty of demand, the higher the implicit reorder point will need to be to insure a certain service level.

It is important to note that the reorder point will increase exponentially with the service level. Incremental increases of service level will become increasingly 'expensive' as the 100% are approached. Desired service levels in spare parts are often high; a seemingly small change from for example 95% to 97% service level can have a significant impact on inventory levels.



In the case of spare parts, a simple and effective approach is the grouping of parts according to their strategic importance. Repair parts for example might be higher ranked than maintenance parts, similarly a part that is critical for operation will be higher ranked than a part that can be substituted etc.

A simple and effective way of optimizing service levels is choosing three groups and assigning different service levels (e.g. Group A = 97% service level, Group B = 95% service level, Group C = 85% service level).

A more sophisticated analysis involves the individual optimization of the service level on an SKU basis. This is done by finding the optimal service level at which an SKU produces the best ROI within the constraints set by budgetary and strategic considerations. The analysis behind this optimization however is not trivial given the non-linear correlation between service level and inventory levels and requires dedicated software support.

Summary

From a mathematical viewpoint, quantile forecasts represent a generalization of the classical notion of forecasts. From a practical viewpoint, quantile forecasts are typically superior more accurate for most business situations where risks associated to over and under estimates of the demand are not symmetric. This is particularly true for spare parts.

Despite radical implications of quantile forecasts for retail, manufacturing and wholesale, quantiles have received little attention in the market so far. The simplest explanation is that support for quantile forecasts was close to nonexistent in the software industry.

What did hold the industry back was most likely not lacking insights in statistics, but rather lacking insights in the profound relationship between quantiles and inventory optimization. An additional aspect might have contributed to this late arrival: Quantile models typically require about 10x more processing power compared to classic forecasting models. Without cloud computing, quantile forecasting is hardly possible.

In our opinion, by solving the problem of forecasting intermittent and sparse demand in spare parts management, quantile technology not only provides a strong performance increase, but also makes classic forecasts plain obsolete.

Authors

Joannes Vermorel

An internationally recognized expert on cloud computing and statistical learning, Joannès Vermorel holds a Master of Science degree from the 'École Normale Supérieure de Paris' (ENS Ulm). In 2010, Joannes won the worldwide Microsoft Windows Azure Partner of the Year Award in recognition for his pioneering work in the cloud. Joannes also teaches a software engineering course at the École Normale Supérieure in Paris and is the founder of Lokad.

Matthias Steinberg

CEO of Lokad, Matthias Steinberg was previously Vice President at Summit Partners LTD, a leading global private equity and venture capital firm. His prior experience includes Airbus Industries and The Boston Consulting Group. He holds a Master of Engineering from RWTH Aachen and an MBA from INSEAD.

Lokad

Lokad is a technology company that focuses on Big Data analytics software for retail networks, wholesale and eCommerce. Client solutions include inventory optimization, (loyalty) marketing optimization, and out-of-shelf monitoring. The company is the winner of the 2010 Microsoft Worldwide Partner of the Year Award and is recognized as an international leader in cloud computing technology.

Microsoft Partner Network

2010 PARTNER OF THE YEAR

Windows Azure Platform

Winner