Demand Driven INSTITUTE
UNITED KINGDOM AND IRELAND

Forget the Forecast Accuracy KPI !

Should we Forecast?
Almost all companies that supply customers 'ex-stock' (eg. CPG, LifeScience etc) have a significant investment in Demand Planning or Forecasting. This usually takes the form of specialist software and people dedicated to generating accurate short term item level forecasts – in monthly or weekly buckets – for the purposes of driving replenishment execution through DRP/MRP/APS. In many cases, the Demand Planning team are 'specialists' and different to the Supply Planning team who use the forecasts to generate the supply plans for execution by Operations.

The key performance metric for Demand Planning is, of course, 'forecast accuracy' and this is usually measured using some form of aggregated forecast error (ie. gross error) as a percentage of the total forecast or demand across a portfolio for a given time bucket (usually monthly) with an appropriate lead-time off-set. Measured in this way, consistent forecast accuracy of 80% is generally considered 'world class' and is extremely rare.

The rationale for setting store by such a 'forecast accuracy' KPI is obvious: in a 'forecast push' replenishment environment, it is the forecast of future demand that is used to drive the supply activities that will ensure the right levels of stock are located in the right locations to meet that demand and achieve the desired level of OTIF service.

Of course 100% forecast accuracy isn't expected as its impossible and, anyway, safety stock is held to buffer against the forecast error.

But as all Demand Planners know, the first questions to be asked when service issues occur are around the forecast and, of course, in some cases extremes of demand can be identified as the cause of a stock out.

So far so good – the process seems entirely sensible: reasonable forecast accuracy can be achieved, safety stocks can be used to compensate for forecast error and so long as Operations make on time what is required, desired service levels will be achieved and stock turns will be roughly as expected given average order quantities?

Right?

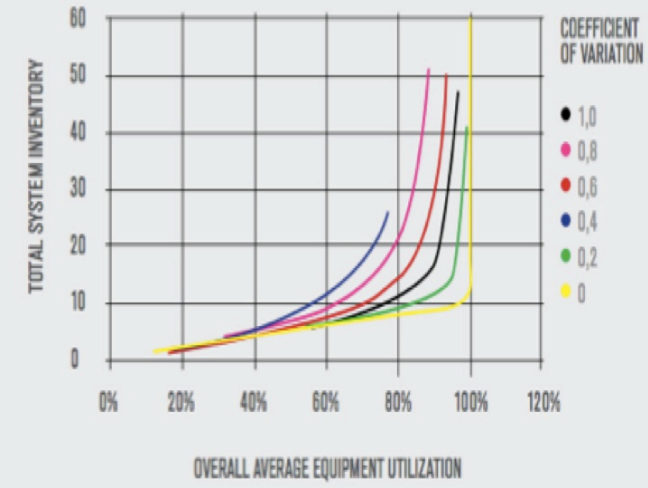
Wrong!!

Why Using Forecasts to Drive Replenishment Doesn't Work
There are two key reasons why desired supply chain performance can't be achieved using this 'forecast push' process and the root cause for both is 'variability':

1 Forecasts – the 80% level of world class portfolio accuracy actually hides a huge amount of inaccuracy. 80% is only

achieved due to the weighting of the relatively few high volume / low variability demand pattern sku's that can enable forecast accuracies of 90%+. Due to 'pareto', the majority of sku's in most portfolios are of medium and high variability for which forecast accuracies of 60% and below are the norm. Most companies spend far too little time on managing their safety stock levels accurately and very rarely use a properly calculated and maintained 'forecast error with offset' algorithm. And even if they did, the entire logic behind DRP/MRP generates exception messages whenever safety stock use is predicted. If the exception messages are responded to the forecast error is simply 'bounced' up the supply chain in the form of supply schedule adjustments that disrupt flow and generate variability. Of course, the exception messages can be ignored but the result is service or inventory issues that have to be responded to sooner or later with the same result. Most companies try to minimise frequent schedule changes by buffering their factories with immensely exaggerated lead-times and / or time fences, which is why so much of a products planning, and actual, lead-time is non-value add.

2 Fixed Lead Times and Dependent Demand – a fundamental assumption of MRP is that lead-times are fixed but this is wrong. Lead-times are directly proportional to levels of variability and increase in an exponential manner when capacity utilisation is high. This can be graphically illustrated as follows:



There is even a mathematical equation that quantifies these relationships and in its most simple form it tells us that:

Average waiting Time in a Queue = Variability (of arrivals & process time) x Utilisation x ave. processing Time (1)

or

Q=VUT

And because MRP uses dependent demand, any extension in a lead-time ripples out across the supply chain and disrupts flow, particularly as, in such an environment, delays accumulate but gains do not.

Variability
As we have seen, inaccurate forecasts blown through MRP result in schedule changes and unplanned change-overs that inevitably impact the planned schedules of other products and their lead-times. Even if the forecasts were 100% accurate, however, natural process variability and sensitivity to load would still lead to fluctuating lead-times with knock on effects that result in congestion and further waves of variability throughout the supply chain – particularly if capacity utilisation is high. The inevitable result is excessive lead times and, through Little's Law, excessive levels of stock, use of unplanned over-time and continuing service issues.

All of this can be understood in the context of what supply chains really are – flows of materials through value add processes that are prone to turn into queues whenever there is variability or lack of flexibility, particularly in the presence of high capacity utilisation. This lack of flexibility can either be buffered with time (make the customer wait), capacity (chop and change the schedule with unplanned change-overs) or finished goods inventory.

If the supply chain isn't flexible enough to respond perfectly to demand then these buffers are inevitable. MRP's erroneous assumption of fixed lead-times and dependent demand are a serious inhibitor of flexibility and, ironically, the frequently seen response of MRP system lead-time inflation to compensate only makes the situation worse – not least by causing more work to be released to the factory floor thereby increasing capacity utilisation and making the problem worse.

In reality, supply chains will exhibit all three buffers and the aggregate buffer is always far more than that necessary because of the variability injected by the forecast error and its amplification by fixed lead-times and dependent demand.

What can be done?
Supply chains will always suffer from variability to some extent, but it can be minimised. One method is implementation of Lean which is:

"…….fundamentally about minimising the cost of buffering variation" (2)

Activities such as TPM, TQM, Standard Work, SMED and batch size reduction all contribute to flexibility and reliability that reduces 'supply side' variation.

Another well known Lean tool is Pull and Demand Driven execution is simply a cross enterprise(s) version of Pull that allows the entire 'end to end' supply chain to flow in line with real demand.

The essence of Demand Driven is the deliberate positioning of multiple planned, but independent, inventory locations up and down the supply chain that are autonomously replenished, in an efficient and stable sequence, to a carefully calculated stock target, in line with demand. The inventories de-couple 'value add' activities thereby preventing residual variability from being propagated and amplified across the supply chain. And because replenishment is driven by demand, not the forecast, it is always correct and schedule interruptions that would otherwise be a source of variability (or need to be buffered) are eliminated.

In a Demand Driven supply chain forecasting is still important for S&OP, Event Management and stock target sizing (eg. if there is trend or seasonality the targets need to reflect it appropriately), but high levels of time phased item level forecast accuracy are no longer needed and that particular KPI can be dispensed with.

Because of the far lower levels of variability in a Demand Driven supply chain, relative to that using 'forecast push', its implementation consistently delivers significant performance improvements due the elimination of cost generating buffers (3):

- Achievement of planned service levels
- Reductions in inventory of 30-50% with significantly shorter lead-times
- Higher capacity utilisation leading to cost reductions of c20%
- Lead-time reductions of up to 85%

And Demand Driven can now be inexpensively simulated, piloted and implemented across large and complex networks using robust, functionality rich and globally tested 'Software as a Service' that simply uploads/downloads FTP files with transaction systems.

To learn more, visit: www.demanddriveninstitute.com

References

1 The Kingman Equation, see Hopp & Spearman 'Factory Physics' 1996
2 Hopp & Spearman, 'To pull or not to pull' M&SOM Vol 6, No 2, 2004, p133-148
3 Case studies can be supplied on request and some can be viewed at: www.demanddrivenworld.com

Simon Eagle 2015