Methods for Intermittent Demand Forecasting

1 Introduction

1.1 What is intermittent demand?

Intermittent demand or ID (also known as sporadic demand) comes about when a product experiences several periods of zero demand. Often in these situations, when demand occurs it is small, and sometimes highly variable in size.

ID is often experienced in industries such as aviation, automotive, defence and manufacturing; it also typically occurs with products nearing the end of their life cycle. Some companies operating in these areas observe ID for over half the products in their inventories. In such situations there is a clear financial incentive to inventory control and retaining proper stock levels, and therefore to forecasting demand for these items.

1.2 Forecasting difficulties

The many zero values in ID time-series render usual forecasting methods difficult to apply. For example, single exponential smoothing (SES), proposed in 1956, was the first forecasting method to be applied to intermittent demand. The forecast of demand in the next period is a weighted average between two quantities, defined:

$$F_{t+1} = \alpha X_t + (1-\alpha)F_t$$

where F_t denotes the forecast for time period t, X_t denotes the actual demand observed in period t, and α is a smoothing parameter which can be adjusted between 0 and 1. Higher α will produce a forecast which is more responsive to recent changes in the data, whilst also being less robust to noise.

Unfortunately, SES is known to perform poorly in forecasting for ID, since there is an upward bias in the forecast in the period directly after a non-zero demand. Can we develop specific forecasting methods for ID that do better? This report explores a few different classes of method, and also discusses some error metrics used to evaluate them.

2 Ad-hoc forecasting methods

The ad-hoc class of forecasting methods are not underpinned by theoretical results. However, it is often the case that such methods can be demonstrated to give good, robust forecasts through empirical experiments and practical use. These methods are, therefore, useful, and are the methods most used in industry today.

2.1 Croston's method

The first ID-specific method was proposed by Croston [1]. His insight was that estimating demand probability (via interval size) and demand size separately was more intuitive and accurate. Let Z_t be the estimate of mean non-zero demand size for time t, V_t the estimate of mean interval size between non-zero demands. X_t again denotes actual demand observed at time t, and q is the current number of consecutive zero-demand periods. Y_t will denote an estimate of mean demand size (ie. taking zero demands into the calculation). Then:

$$\text{If } X_t \neq 0 \text{ then } \begin{cases} Z_{t+1} &= \alpha X_t + (1-\alpha)Z_t \\ V_{t+1} &= \alpha q + (1-\alpha)V_t \\ Y_{t+1} &= \frac{Z_{t+1}}{V_{t+1}} \end{cases}$$

$$\text{If } X_t = 0 \text{ then } \begin{cases} Z_{t+1} &= Z_t \\ V_{t+1} &= V_t \\ Y_{t+1} &= Y_t \end{cases}$$

Croston showed that this process gave significantly better forecasts than SES for some ID data.

There are limitations to this method; the first is bias. Croston argued that for such estimates V_t and Z_t, Y_{t+1} as a forecast for the demand next period would be unbiased. However, Syntetos and Boylan [6] showed that, since $E[\bar{X}_t] = E[\frac{Z_t}{V_t}] \neq E[Z_t]\frac{1}{E[V_t]}$, bias is, in fact, present. Additional drawbacks include the lack of independent smoothing parameters for demand size and interval size, the assumption that demand size and demand interval are independent (this is generally too strong), and there is no way to deal with product obsolescence.

2.2 Adjusted Croston methods

Many adaptations of Croston's method have been suggested to deal with some of the aforementioned issues. In [6], the authors propose an adjustment, known as the Syntetos-Boylan Approximation (SBA), to Croston's forecast Y_t, namely that it should be multiplied by a factor of $(1 - \frac{\alpha}{2})$, and claim that the new forecast will be approximately unbiased, since

$$\begin{aligned} E\left[\left(1-\frac{\alpha}{2}\right)\left(\frac{Z_t}{V_t}\right)\right] &= \left(1-\frac{\alpha}{2}\right) E\left[\frac{Z_t}{V_t}\right] \\ &= \left(1-\frac{\alpha}{2}\right)\left(\frac{\mu}{p} + \frac{1}{2}\frac{\partial^2(\frac{\mu}{p})}{\partial p^2}\ \mathrm{Var}(p)\right) \\ &= \left(\frac{2-\alpha}{2}\right)\left(\frac{\mu}{p} + \frac{\alpha}{2-\alpha}\mu\frac{p-1}{p^2}\right) \\ &= \frac{\mu}{p}\left(\frac{2-\alpha}{2} + \frac{\alpha}{2}\frac{p-1}{p}\right) \approx \frac{\mu}{p} \end{aligned}$$

A method incorporating the possibility of item obsolescence (where an item is no longer demanded at all) is offered by Teunter et al. [7]. It differs from Croston's method and SBA in that it estimates the probability of non-zero demand (rather than interval size), and in that the estimates are updated every period, rather than just when demand occurs. Let D_t be an indicator of a non-zero demand at time t, P_t the estimate of demand for time period t. The Teunter, Syntetos and Babai (TSB) method is:

$$\text{If } D_t = 1 \text{ then } \begin{cases} P_{t+1} &= \beta(1) + (1-\beta)P_t \\ Z_{t+1} &= \alpha X_t + (1-\alpha)Z_t \\ Y_{t+1} &= P_{t+1}Z_{t+1} \end{cases}$$

$$\text{If } D_t = 0 \text{ then } \begin{cases} P_{t+1} &= (1-\beta)P_t \\ Z_{t+1} &= Z_t \\ Y_{t+1} &= P_{t+1}Z_{t+1} \end{cases}$$

Note that different smoothing parameters α and β are used here for smoothing demand size and demand probability.

The key advantage of this method is that updating the forecasts at each time period (whether demand occurs or not) allows the estimate P_t to approach zero if there is a long run of periods without demand. By contrast, the interval estimate calculated by Croston's would remain unchanged. In a practical setting, TSB allows decisions to be taken over whether to continue to stock items or not. One idea might be to set some threshold for p, such that if it is exceeded (just once or for a number of consecutive periods) it would be decided that the product is now obsolete.

A second, different method dealing with obsolescence was proposed recently by Prestwich et al. [5] and is known as Hyperbolic Exponential Smoothing (HES). The method combines the Croston method with a Bayesian approach to derive a new forecast, where if Z_t and V_t are the smoothed estimates of demand size and interval length, and T_t is the current number of periods since a demand was last seen, then

$$F_t = \frac{Z_t}{V_t + \beta\frac{T_t}{2}}$$

is the forecast for time t. Note β is the smoothing parameter for interval length. The main difference between TSB and HES is that HES decays hyperbolically over a series of zeros, whereas TSB decays only exponentially.

3 Model-based forecasting methods

A wholly different approach is that of using statistical models to model intermittent demand time series. The idea of using such models is promising, a large part of the attraction being that we can back up our models with theoretical results, in contrast to the ad-hoc class. Developing statistical models for ID forecasting may also prove to aid understanding of the properties which govern ID time-series.

3.1 ARMA models

Auto-regressive moving average (ARMA) models are used to describe stationary stochastic processes. An ARMA(p, q) model comprises two parts:

- an AR(p) process $X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t$, where c is a constant, φ_i are parameters of the model, ϵ_t is random noise.
- an MA(q) process $X_t = \mu + \sum_{i=1}^{q} \theta_i \epsilon_{t-i} + \epsilon_t$, where μ is a constant, θ_i are parameters.

which combine to give our full ARMA(p, q) model:

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \varphi_i X_{t-i} + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

In other words we model the points in our series as being dependent on the previous p points (auto-regressive) and on the previous q residuals (moving-average). For intermittent demand, however, an ARMA process is an inappropriate model since it allows values that aren't non-negative integers. We need an adjustment which adds this constraint. Two ideas are quite commonly used.

3.2 DARMA models

Discrete ARMA (DARMA) models take values on a discrete set, according to the results of some random variables. These processes are best illustrated by example, and a simple one is the DAR(1) model. If Y_t here is the t-th value of the series, $\{V_t\}$ are independent Bernoulli(α) random variables, and $\{Z_t\}$ are i.i.d random variables defined on the set $\{0,1,2,\dots\}$:

$$Y_t = V_t Y_{t-1} + (1 - V_t) Z_t$$

In other words, the next value in the series is not an SES-style weighted average of two quantities (the previous value or the value of a discrete-valued random variable), but rather a random choice between the two. A full DARMA(p, q) model is a random choice between autoregressive and moving-average terms.

3.3 INARMA models

Integer-valued ARMA (INARMA) models take previous autoregressive and moving-average terms, 'thin' them, and add some random non-negative integer. The INARMA(p, q) process is given by:

$$Y_t = \sum_{i=1}^{p} \alpha_i \circ Y_{t-i} + \sum_{i=1}^{q} \beta_i \circ Z_{t-i} + Z_t$$

where the thinning operator $\circ$ is defined by

$$\beta \circ Z = \sum_{i=1}^{Z} X_i, \qquad X_i \text{ is Bernoulli}(\alpha)$$

INARMA models for forecasting ID were investigated in [3]. Once parameters for the model have been selected, one way to get a forecast is to find the conditional expected value of the process, which is based on minimising mean squared error.

4 Alternative methods

Many alternative methods have been proposed; two of these are displayed here.

4.1 Bootstrapping

Bootstrapping is a statistical technique involving random sampling with replacement. In 2004 Willemain et al. [8] proposed a method using bootstrapping on previous observations of non-zero demand to forecast demand over some lead time (the interval between replenishment ordering and arrival).

Two key ideas are present in the method. Firstly, to avoid making forecasts that can only take the same values as have occurred previously, a jittering process is used, allowing for more variation. Let Y be the selected previous value of demand, and Z be a standard normal random variable. The jittering process is:

$$Y_{jittered} = 1 + INT\{Y + Z\sqrt{Y}\}$$

with the further constraint that if $Y_{jittered} \leq 0,$ then $Y_{jittered} = Y$. This process is ad-hoc, but is considered to give reasonable results by the authors of [8].

Secondly, to model autocorrelation that might be present in the demand, a two-stage Markov Chain model is used, with the states corresponding to zero and non-zero demand observations. The idea is to forecast a series of zero and non-zero demands first, with the chance of seeing either in the next step dependent on the transition matrix probabilities, which are to be estimated.

The full bootstrap approach as described in [8] goes as follows:

- With the historical data, estimate the transition probabilities for the two-state Markov Chain.
- Generate a sequence of events from the Markov Chain over the desired lead time.
- Replace each non-zero event with a historical value of demand, then jitter these as above.
- Sum the forecast values over the horizon, to gain one estimate of lead-time demand (LTD).

Results shown in [8] demonstrate that the bootstrap method improves significantly upon both Croston's method and SES at forecasting an entire distribution of LTD, over a selection of lead times. There are situations in which having such a distribution to base decisions on, rather than just a point estimate of LTD, would be advantageous.

4.2 Temporal aggregation

The basic idea of temporal aggregation is to combine time periods into blocks. This has the advantage that it could remove zeros in the series, but also the disadvantage that the number of historical observations is greatly reduced.

One recent approach using temporal aggregation for ID was proposed in 2010 by Nikolopoulos et al. [4], who call their method an *aggregate-disaggregate intermittent demand approach* (ADIDA).

There are 3 main stages. The first stage is deciding on a type of aggregation. This includes choosing the number of individual observations to combine in each block, and also whether to aggregate overlapping or non-overlapping blocks of periods. The second stage is to forecast the next value in the aggregated series, which can be done by normal forecasting methods eg. SES, or even the Naïve method (forecasting the next period to take exactly the same value as the last). Finally, the forecast should be disaggregated, or broken down, into time periods of the original size. This can be done using a number of weightings; weights based on the ratio of previous observations in each block could be used, or simply equal weights.

Some empirical evaluations of the method were conducted in [4]. The ADIDA process was found to improve upon both SBA and Naïve forecasting as opposed to applying those methods with no temporal aggregation. Optimal aggregation levels were also investigated; it was unclear that any one level might be optimal in all situations, although setting the level equal to the lead-time plus one was suggested as a good heuristic.

5 Evaluation measures

The performance of any forecasting method needs to be evaluated by some metric, to measure how closely the forecasted value matches the true value. Intermittent demand series turn out to be unusually tricky to evaluate; typical forecasting accuracy metrics are often either inappropriate or even impossible to apply. The metrics presented here are classed as described by Hyndman [2], and some measures that directly link into inventory control are also displayed.

5.1 Scale-dependent metrics

Scale-dependent metrics work with the errors $e_t = X_t - F_t$, ie. the raw difference between the observed demand and the forecasted value. These methods are often easy to compute. Example scale-dependent metrics include the Mean Squared Error (MSE), defined as $\frac{1}{n}\sum_{i=1}^{n}(A_i - F_i)^2$, and the Mean Absolute Error (MAE), defined as $\frac{1}{n}\sum_{i=1}^{n}|A_i - F_i|$; both of these are widely used in traditional forecasting.

However, these methods have disadvantages. The scale-dependency means that comparing the MSE or MAE of multiple time-series is meaningless, which is a serious disadvantage. In addition, if MAE is the only measure to be minimised, then for highly intermittent demand it can often be optimal simply to forecast zero for every period. This is of course no use in a practical inventory control setting.

There is also the Geometric Mean Absolute Error (GMAE), defined as $\sqrt[n]{\prod_{i=1}^{n}|e_t|}$, the geometric mean of the absolute errors, which is specifically recommended by Syntetos and Boylan [6] in some intermittent demand scenarios. This method has the flaw that any error term equal to zero will send the GMAE to zero, and in an intermittent case, this could happen under some methods which are capable of giving zero forecasts (such as the naïve method).

5.2 Percentage-error metrics

Percentage errors, $p_t = 100\frac{e_t}{X_t}$, measure the error for each period as a percentage of the period's observed demand. The removal of scale-dependency allows for comparison of forecasting methods across multiple data series. An example is the Mean Absolute Percentage Error (MAPE) which is simply defined as the mean of $|p_t|$. In an intermittent demand setting, however, X_t is often zero, which would give undefined values of p_t. This alone disqualifies MAPE for use in this setting.

There is also the symmetric MAPE, or sMAPE, defined as the mean of the quantity $200\frac{|Y_t - F_t|}{Y_t + F_t}$. When Y_t is zero here, however, the sMAPE returns a value for that data point which is independent of the forecast; clearly inappropriate.

5.3 Relative errors

We define the relative error as $r_t = \frac{e_t}{E_t}$, where E_t is the error obtained from some other chosen method, known as the 'benchmark' or 'control' method. The idea will be to compare the performance of the new method against this benchmark to get some measure on how much it improves upon that method. Examples include the Median Relative Absolute Error defined by median($|r_t|$), and the Geometric Mean Relative Absolute Error (GMRAE), which simply takes geometric mean instead of the median.

The main issue with these tends to be the benchmark method. The natural choice is the naïve method, but for intermittent demand this means E_t can often be zero, making r_t undefined. It is possible to choose other methods for the benchmark.

5.4 Scale-free errors

The key metric in this category is Mean Absolute Scaled Error (MASE). This is defined:

$$MASE = \frac{1}{n}\sum_{i=1}^{n}\frac{|A_t - F_t|}{\frac{1}{n-1}\sum_{i=2}^{n}|Y_i - Y_{i-1}|}$$

where the denominator can be thought of as the MAE for the naïve method (since $F_t = Y_{i-1}$). Hyndman (who proposed this measure) recommends it in [2] as the measure to use when studying intermittent demand due to its

robustness and versatility; MASE is never undefined or infinite for non-trivial cases. However, it seems to suffer the same problem as MAE in an ID setting; a zero forecast may well prove 'best'.

5.5 Inventory-related measures

For demand-related scenarios, forecast error measures relating to inventory performance can also be applied. These include

- Service level α - measures the probability that demand will be below stock level if we replenish stock based on our forecasts (can be thought of as the probability an arbitrary customer is served straight away).
- Service level β - measures the quantity of demand met immediately over the total demand (or the probability an arbitrary unit of demand is satisfied straight away).
- Periods In Stock (PIS) - the formula for this is defined $PIS_n = -\sum_{i=1}^{n}\sum_{j=1}^{i}(A_j - F_j)$. The idea is that we assume the existence of a 'fictitious stock' to which the number in our forecast at each time period is delivered, and then the number of units demanded in that time period are removed. A small example - let us imagine a case where $n = 2$. Suppose the forecast for the first period was 2 units, and the forecast for the second was 1 unit. Also suppose the actual demand was 0 in both periods. Then the first two units will be delivered to the fictitious stock in period 1, and since neither are demanded, they both spend 2 periods in stock. Another unit will be delivered in period 2; it was not demanded, so spends one period in stock. The total PIS in this case therefore is 5.

6 Conclusions

The limitations of usual forecasting methods, such as SES, in the intermittent demand case has prompted a number of different ID-specific approaches to forecasting. The first method developed was Croston's; its aptitude for ID forecasting was examined, as well as its drawbacks. Methods that attempted to address some of these drawbacks were considered. More theoretical, model-based approaches were introduced as a promising alternative avenue, as well as methods based in bootstrapping and temporal aggregation. Finally, examples of error metrics that can be used to measure accuracy of forecasts were considered , along with their suitability in the ID setting.

References

[1] J. D. Croston. Forecasting and stock control for intermittent demands. *Operational Research Quarterly (1970-1977)*, 23(3):pp. 289–303, 1972.

[2] Rob Hyndman. Another look at forecast accuracy metrics for intermittent demand. *Foresight: the International Journal of Applied Forecasting*, 4(4):43–46, 2006.

[3] M. Mohammadipour and Brunel University. Intermittent demand forecasting with integer autoregressive moving average models. *PhD Thesis*, 2009.

[4] K. Nikolopoulos, A. A. Syntetos, J. E. Boylan, Fotios Petropoulos, and V. Assimakopoulos. An aggregate-disaggregate intermittent demand approach (adida) to forecasting : an empirical proposition and analysis. *Journal of the Operational Research Society : OR*, 62(3, (3)):544–554, 2011.

[5] S.D. Prestwich, S.A. Tarim, R. Rossi, and B. Hnich. Forecasting intermittent demand by hyperbolic-exponential smoothing. *International Journal of Forecasting*, 30(4):928 – 933, 2014.

[6] Aris A. Syntetos and John E. Boylan. The accuracy of intermittent demand estimates. *International Journal of Forecasting*, 21(2):303 – 314, 2005.

[7] Ruud H. Teunter, Aris A. Syntetos, and M. Zied Babai. Intermittent demand: Linking forecasting to inventory obsolescence. *European Journal of Operational Research*, 214(3):606 – 615, 2011.

[8] Thomas R. Willemain, Charles N. Smart, and Henry F. Schwarz. A new approach to forecasting intermittent demand for service parts inventories. *International Journal of Forecasting*, 20(3):375 – 387, 2004.