Intermittent Demand Forecasts with Neural Networks

Nikolaos Kourentzes*

Lancaster University Management School, Lancaster, LA1 4YX, UK

Abstract

Intermittent demand appears when demand events occur only sporadically. Typically such time series have few observations making intermittent demand forecasting challenging. Forecast errors can be costly in terms of unmet demand or obsolescent stock. Intermittent demand forecasting has been addressed using established forecasting methods, including simple moving averages, exponential smoothing and Croston's method with its variants. This study proposes a neural network (NN) methodology to forecast intermittent time series. These NNs are used to provide dynamic demand rate forecasts, which do not assume constant demand rate in the future and can capture interactions between the non-zero demand and the inter-arrival rate of demand events, overcoming the limitations of Croston's method. In order to mitigate the issue of limited fitting sample, which is common in intermittent demand, the proposed models use regularised training and median ensembles over multiple training initialisations to produce robust forecasts. The NNs are evaluated against established benchmarks using both forecasting accuracy and inventory metrics. The findings of forecasting and inventory metrics are conflicting. While NNs achieved poor forecasting accuracy and bias, all NN variants achieved higher service levels than the best performing Croston's method variant, without requiring analogous increases in stock holding volume. Therefore, NNs are found to be effective for intermittent demand applications. This study provides further arguments and evidence against the use of conventional forecasting accuracy metrics to evaluate forecasting methods for intermittent demand, concluding that attention to inventory metrics is desirable.

*Correspondance: N Kourentzes, Department of Management Science, Lancaster University Management School, Lancaster, Lancashire, LA1 4YX, UK. Tel.: +44-1524-592911

Email address: n.kourentzes@lancaster.ac.uk (Nikolaos Kourentzes)

Preprint submitted to International Journal of Production Economics January 10, 2013

Keywords: Intermittent demand, Neural Networks, Croston's Method, Forecasting, Slow moving items

1. Introduction

There are many cases where items in an inventory control system are requested infrequently, resulting in sporadic or intermittent demand. Time series of intermittent demand differ from conventional series in the respect that they have multiple periods of zero demand (Croston, 1972). Willemain et al. (2004) identified intermittent demand in several scenarios, such as heavy machinery and respective spare parts, aircraft service parts, electronics, maritime spare parts, etc, while Syntetos and Boylan (2005) explored intermittency in the automotive spare parts. Ghobbar and Friend (2003) looked at the demand of aircraft maintenance parts, which are often of high value. Johnston et al. (2003) identified that such items can account for up to 60% of the total stock value. Due to their slow moving nature, such items are at greatest risk of obsolescence. This can have substantial impact on the operations of organisations, which tie resources in stocking items of this nature. In practice, as Ghobbar and Friend (2003) observed, companies hold more stock than needed due to inaccurate demand expectations and often without achieving the desired service levels. Inaccuracies can arise from both the magnitude and the timing of the demand. Therefore, in order to support inventory holding and replenishment decisions accurate demand forecasts are necessary.

Croston (1972) observed that conventional time series methods, such as exponential smoothing, did not perform well and proposed an alternative forecasting method suited to intermittent demand time series. Research has shown that Croston's method is appropriate to use for forecasting intermittent demand time series and provides improvements over conventional time series methods, for e.g. see Willemain et al. (2004) and Johnston and Boylan (1996). Syntetos and Boylan (2001) showed that the original Croston's method is biased and proposed a modified version that corrected the problem (Syntetos and Boylan, 2005), demonstrating improved accuracy. Levén and Segerstedt (2004) proposed an alternative modification to Croston's method, attempting to avoid the bias of the original method. Boylan and Syntetos (2007) showed that this is not the case and the method by Levén and Segerstedt (2004) is still biased, however in a different manner. Its bias, in contrast

to Croston's method, does not vary with the smoothing of the demand intervals. They demonstrated that this method is more biased than Croston's, in particular for highly intermittent series.

Hyndman and Shenstone (2005) argue that Croston's method is an ad-hoc method with no properly formulated underlying stochastic model. As it is based on exponential smoothing, they argue that it assumes continuous data, including negative values, which of course is not true for intermittent demand data that are integer-valued and non-negative. Furthermore, Croston (1972) stated that his method assumes that the demand size and the inter-demand intervals are independent. However, Willemain et al. (1994) questioned this assumption of independence. Nonetheless, this has been retained in later work that improves upon Croston's original method, see for e.g. Syntetos and Boylan (2001, 2005); however this has not been proven to be true and in many cases it is taken for granted without testing this assumption on the available data. Therefore, as Hyndman and Shenstone (2005) find, Croston's method is inconsistent with the properties of intermittent demand, but of practical usefulness as it has been shown empirically to outperform conventional methods (Willemain et al., 1994; Syntetos and Boylan, 2001).

This paper proposes a Neural Network (NN) method for intermittent demand time series forecasting. The motivation behind this work is based on the nonparametric, data assumption free nature of NNs. In particular, the feedforward multilayer perceptrons, that are employed here, have been proven by Hornik et al. (1989) and Hornik (1991) to be universal approximators, therefore, in theory, able to capture the data generating process of intermittent demand time series. Zhang et al. (1998) and Dahl and Hylleberg (2004) argue that NNs are flexible models that do not require human experts to prescribe rigid model structures. The proposed NN method allows for interaction between the demand size and the inter-demand intervals of demand events, or their lags, if such can be identified from the data and there is no need for expert input. Naturally their flexible nature is advantageous for capturing the intermittent demand structure. Gutierrez et al. (2008) proposed a NN method for lumpy demand time series that outperformed, in their experiments, Croston's method and Syntetos and Boylan (2005) modification. However, as section 2 discusses, it has significant limitations. The work in this paper aims to provide a less restrictive framework. The basic concepts from Croston's method are retained, but expanded to allow for modelling dynamics in either the demand volume or its inter-demand

intervals and any interactions between them. Boylan and Syntetos (2007) observed that little work has been done to model this interaction, even though their independence has been questioned (Willemain et al., 1994). Furthermore, the proposed method expands on previous work by providing dynamic forecasts, allowing the predicted demand rate to vary for different forecast horizons, thus being more flexible than the constant that is the norm in intermittent demand forecasting. NNs have often been seen as data hungry models. Markham and Rakes (1998) provide evidence that NNs require large sample size to outperform conventional statistical methods. Intermittent demand time series, especially when only non-zero demand is modelled, as is the case here, have typically very few observations. Therefore, conventional NN training is not practical and regularised networks are proposed instead to mitigate the problem of small sample size.

The proposed NNs for intermittent demand are evaluated against a number of fast moving and intermittent demand forecasting benchmark models. These are assessed using both bias and accuracy metrics of point forecasts, which is the conventional approach for comparing different forecasting methods, as in Syntetos and Boylan (2005). Gardner (1990), Teunter and Duncan (2009) and Syntetos et al. (2010) dispute this and argue in favour of stock control metrics, which are more directly related to the decision making problems that an organisation faces. Assessing the performance of the proposed method and benchmarks in both ways provides evidence that relying solely on forecasting metrics can result in misleading findings. Although no superiority of the NNs over benchmark methods is identified when forecasting accuracy is considered, a clear advantage of NNs is found when looking at stock control metrics. Service level improves significantly, without increasing the holding volume substantially. Therefore, the proposed NNs offer important improvements over the standard benchmarks.

The rest of the paper is structured as follows: section 2 discusses existing forecasting approaches for intermittent demand items and presents the novel NNs, while sections 3 and 4 discuss the experimental setup and the results respectively. Section 5 concludes after a discussion of the findings.

2. Methods

2.1. Croston's Method and Modifications

The standard method for intermittent demand forecasting is considered to be Croston's method, as proposed originally by Croston (1972) and later

corrected by Rao (1973). Instead of forecasting an item in the conventional way, for instance using Exponential Smoothing, the time series is broken into its constituent elements; the non-zero demand size z_t and the inter-demand intervals x_t. Note that the original time series contains periods of zero demand. Demand is assumed to occur as a Bernoulli process, therefore the inter-demand intervals are geometrically distributed with a mean $\bar{x}$. The non-zero demand is assumed to follow the normal distribution. Both z_t and x_t are forecasted using single exponential smoothing (SES), with the smoothing parameter α identical for both. Croston (1972) suggests a smoothing parameter between 0.1 and 0.3, while Syntetos and Boylan (2001) advise α to be no greater than 0.15. The resulting estimates z'_t and x'_t are only updated when demand occurs and remain constant otherwise. The forecast Y'_t is given by:

$$Y'_t = \frac{z'_t}{x'_t}. \tag{1}$$

The multi-step ahead forecast is a constant with value equal to Y'_t. Note that if demand occurs every period, then x_t is a vector of ones and therefore $Y'_t = z'_t$, in other words Croston's method is identical to SES. This method will be named $CR - SES$ hereafter. Syntetos and Boylan (2001) showed that Croston's method is biased, due to the division in (1), and suggested a modified version (Syntetos and Boylan, 2005):

$$Y'_t = \left(1 - \frac{\alpha}{2}\right) \frac{z'_t}{x'_t}. \tag{2}$$

This modified version uses the same assumptions as the original method, and in this work is named $CR - SES - SB$. As discussed in section 1 these assumptions have been challenged. However the modified version has demonstrated good empirical performance, superior to the original method (Syntetos and Boylan, 2005, 2006). Shale et al. (2006) showed that if the orders arrive as a Poisson process then the appropriate modification becomes:

$$Y'_t = \left(1 - \frac{\alpha}{2 - \alpha}\right) \frac{z'_t}{x'_t}. \tag{3}$$

This model is named $CR - SES - SH$ in this paper. The same study discusses an obvious modification of Croston's method, which replaces SES for the estimation of z'_t and x'_t in 1 with Moving Averages (MA) of length k,

resulting in method $CR-MA$. The inversion bias discussed by Syntetos and Boylan (2001) exists again and the respective modifications for the Syntetos-Boylan approximation is:

$$Y'_t = \left(\frac{k}{k+1}\right)\frac{z'_t}{x'_t}, \tag{4}$$

resulting in model $CR-MA-SB$ and for Shale et al. is:

$$Y'_t = \left(\frac{k-1}{k}\right)\frac{z'_t}{x'_t}, \tag{5}$$

which is named here $CR-MA-SH$. A basic benchmark model that follows Croston's method philosophy, and does not appear in the literature, would be to assume no smoothing in either z_t and x_t, essentially estimating both z'_t and x'_t using the random walk, i.e. their forecast would be equal to the last historical value. This is equivalent to setting either α or k in the above models to 1. This parameterless benchmark is employed in this study to show the improvement in the performance of the above models, if any, due to forecasting the demand and the inter-demand intervals using either exponential smoothing or moving averages, which require setting a parameter. This model is named $CR-Naive$.

Careful consideration of equation 1 reveals that Croston's method does not provide a demand forecast for the time series, rather than a "demand-rate" forecast. Y'_t in this case is the average expected demand in each future period and not the point forecast of the demand for the future periods. This is a significant difference in comparison to conventional time series models. For example if the forecast is 0.25 that should be interpreted that there is a demand of 1 unit over four periods, or the demand-rate per period is 0.25. Therefore, the model output is not violating the integer valued nature of intermittent demand time series.

2.2. Conventional Time Series Models

Syntetos and Boylan (2005) report that Moving Averages (MA) and Single Exponential Smoothing (SES) are often used in practice to forecast intermittent demand time series. Given demand d_t under MA of order k the forecast is:

$$Y'_t = \frac{\sum_{i=1}^{k} d_{t-i}}{k}. \tag{6}$$

For SES with smoothing parameter $0 \leq \alpha \leq 1$ the forecast is:

$$Y'_t = \alpha d_{t-1} + (1 - \alpha) Y_{t-1}. \tag{7}$$

Due to the periods of zero demand both models are expected to perform badly for a time series that exhibits intermittent demand. A simple benchmark that any more complex forecasting model should outperform is the random walk, or naive model, which essentially assumes that the forecast is equal to the last observed value. Both MA and SES devolve into the $Naive$ by setting either k or α to 1 respectively.

2.3. Neural Networks

Neural Networks (NNs) are flexible nonlinear data driven models that have attractive properties for forecasting. Traditional statistical time-series methods can often fit poorly to the underlying data generating process, because of their restrictive functional forms. On the other hand, NNs are flexible models that learn the data generating process from the data without requiring human intervention. Furthermore, traditional forecasting methods often rely on restrictive data assumptions, which the neural networks do not have. Zhang et al. (1998) list multiple forecasting applications where NNs have been employed successfully. Adya and Collopy (1998) found that of 73% of the papers reviewed NNs to outperform established benchmarks. The most commonly used form of NNs for forecasting is the feedforward Multilayer Perceptron (MLP). Zhang et al. (1998) provide a detailed description of these models and how to use them for forecasting. The one-step ahead forecast Y'_t is computed using inputs that are lagged observations of the time series. I denotes the number of input p_i of the NN. The functional form is

$$Y'_t = \beta_0 + \sum_{h=1}^{H} \beta_h g \left(\gamma_{0i} + \sum_{i=1}^{I} \gamma_{hi} p_i \right). \tag{8}$$

In equation (8), $\mathbf{w} = (\boldsymbol{\beta}, \boldsymbol{\gamma})$ are the network weights with $\boldsymbol{\beta} = [\beta_1, \ldots, \beta_H]$ and $\boldsymbol{\gamma} = [\gamma_{11}, \ldots, \gamma_{HI}]$ for the output and the hidden layers respectively. The β_0 and γ_{0i} are the biases of each neuron, which function as the intercept in a regression for each neuron. H is the number of hidden nodes in the network and $g(\cdot)$ is a non-linear transfer function, which is usually either the sigmoid logistic or the hyperbolic tangent (TanH) function and provides the nonlinear capabilities to the model. Their functional form allows them to

model interactions between inputs, if any. The hidden nodes are connected to a linear output node that produces the forecast. The network can be seen as a nonlinear autoregressive model that can be naturally extended to include multivariate inputs (Connor et al., 1994).

Neural networks are often seen as "data-hungry" models, requiring large samples to train on. This can be a major drawback for intermittent demand applications. The reason for this is their large number of degrees of freedom; for a typical three layer MLP it is $(I + 1)H + 1$. A large sample is also needed to accommodate the complex nonlinear optimisation problem that needs to be solved to find the connecting weights and biases. In order to avoid overfitting to the in-sample data it is common practice with NNs to split the in-sample data into two subsets, the training and the validation samples. The network is optimised on the training sample, while its accuracy is recorded on the validation sample. If the error on the validation sample starts increasing, while on training is reducing, that signals that the network is overfitting to the data and training is stopped. The weights and biases that give the lowest error on the validation sample are retained. It is crucial to avoid overfitting in order to achieve good out-of-sample forecasting performance.

Therefore, only a portion of the in-sample data is available to solve a hard nonlinear optimisation problem with many degrees of freedom. Training can often get trapped in local minima of the error surface. In order to avoid a poor quality local minimum, training is initialised several times with different random starting weights and biases. The best initialisation is retained as the final model, which is selected on minimum validation sample error.

NNs have been explored by Gutierrez et al. (2008) for lumpy demand forecasting applications. Lumpy demand time series exhibit high variability of the non-zero demand (Syntetos et al., 2005). They use standard $MLPs$ with a particular architecture and set of inputs. They propose using 3 hidden nodes in a single hidden layer with two inputs. The first input is the last observed demand. Note that this is not the last non-zero demand as in the conventional Croston's method. The second input is the number of periods separating the last two non-zero demand transactions. The output of the model represents the predicted demand, in a similar way to SES. The models are trained using the standard back-propagation algorithm with momentum (Rumelhart et al., 1988). They report that NNs outperformed SES, $CR - SES$ and $CR - SES - SB$ with different smoothing parameters on a set of 24 time series.

It is interesting to focus on the NN training particulars of this study. The

time series contained 967 daily observations, providing a substantial sample for NNs to train effectively. Gutierrez et al. (2008) do not use a validation sample that is useful in mitigating the tendency of NNs to overfit in the training data (Bishop, 1996), as discussed above. Therefore, they provide strong evidence that NNs can perform well in intermittent demand applications even with very rudimentary training setups. However, as they identify, this requires abundance of data, which is not seen in other intermittent demand studies that use only a few years of monthly data, for e.g. see Syntetos and Boylan (2005).

In this paper the implementation of NNs for intermittent demand is revisited in order to provide a more flexible framework. Initially the sample size problem is addressed, so as to allow such models to be applicable in a variety of intermittent demand problems, in particular for short time series. Similar to traditional statistical models, NNs are optimised based on a one-step ahead mean squared error loss function. As Hornik et al. (1989) showed neural networks are universal approximators and in practice prone to overfitting, unless special care in training is taken. This can be mitigated by using regularised cost functions for neural network. Bishop (1996) discusses a simple and commonly used form of regularisation, where the cost function F is changed into:

$$F = \gamma \frac{\sum_{i=1}^{N} (e_i)^2}{N} + (1-\gamma) \frac{\sum_{j=1}^{n} (w_j)^2}{n}. \tag{9}$$

The first part of the cost function is the conventional mean squared error loss of N one-step ahead e_i errors. The second part of F is keeping the weights of the network w_j small, by penalising large weights, essentially making the network response smoother and less likely to overfit. The performance ratio γ controls the size of regularisation. MacKay (1992) proposes a Bayesian regularisation framework where γ is determined by the data automatically. A major advantage of regularisation is that a validation sample is no longer required in order to avoid overfitting to the in-sample data, thus the implementation of NNs for intermittent demand problems becomes feasible for even small samples.

A straightforward use of NNs in forecasting intermittent demand would be to replace the SES in Croston's method with neural networks. The forecast of non-zero demand z'_t and inter-demand intervals x'_t could be calculated using two separate networks, which would be consecutively divided as in the conventional Croston's method. However such an approach would suffer from

a number of limitations: firstly, it is based on the original Croston's method and therefore is bound to output biased forecasts, due to the division in (1). Secondly, such a model would not allow capturing potential interactions between z_t and x_t. This can be solved easily by forecasting both from a single NN. The neural network now uses lags from both the non-zero demand and inter-demand intervals as inputs and outputs predictions for both separately, capturing any existing interactions. In contrast to the intermittent demand literature, this model does not assume that the demand and the inter-demand intervals are independent and aims to capture any bivariate effects, if present. Moreover, the NNs output a dynamic forecast due to their autoregressive nature. They are able to predict different values for different forecast horizons according to the time series dynamics, in contrast to Croston's method.

The resulting z'_t and x'_t have to be divided as in (1) to provide the forecasts, thus the model suffers from the inversion bias discussed by Syntetos and Boylan (2001). In order not avoid imposing any assumptions, a data driven de-biasing coefficient c is calculated. If Y_t is the in-sample forecast of the neural network, c is calculated by solving the regression:

$$\frac{z_t}{x_t} = cY_t. \tag{10}$$

Multiplying the out-of-sample demand-rate forecasts Y'_t by c de-biases the forecasts. The resulting forecast, in contrast to Croston's method is not constant, but dynamic. The resulting model,[1] $NN - Dual$, is illustrated in figure 1.

A more elegant way to avoid biasing the forecasts is to model the network to output the required demand-rate directly, instead of introducing the subsequent division step. Similar to $NN - Dual$ this model uses lags from both z_t and x_t to output directly Y'_t. The division and the required de-biasing is left to the network to approximate from the data resulting in even greater flexibility. Note that this is different to the model proposed by Gutierrez et al. (2008). The proposed model outputs demand-rate instead of demand point forecasts, making it more suited to the nature of intermittent demand

[1]Simpler NN implementations discussed above, i.e. separate NNs for forecasting z_t and x_t, non-debiased versions and NNs that provide contant forecasts instead of dynamic have been experimented with. It was found that all these models were encompassed by $NN - Dual$ and therefore not presented here.

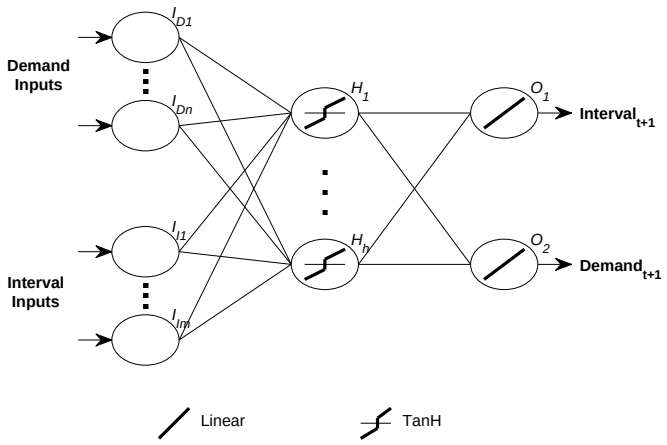

Figure 1: The $NN-Dual$ architecture, with variable number of demand I_{Dn} and interval I_{Im} lagged inputs and hidden nodes H_h, which use the TanH activation function. Two linear output nodes provide the demand and interval forecasts.

time series. These networks do not require any special modelling for bias as is needed by the original Croston's method and its modifications, as the division in (1) is not required. Furthermore, this specification allows for multiple lags of z_t and x_t, therefore capturing any time series dynamics. Again a dynamic forecast is produced.[2] Figure 2 illustrates this class of proposed networks, $NN-Rate$.

3. Empirical evaluation

3.1. Dataset

In order to evaluate the performance of the proposed neural networks a large scale simulation of 1000 items is designed. The dataset used by Syntetos and Boylan (2005) is used to identify realistic parameters to simulate intermittent demand time series. This dataset contains 3000 time series describing the demand of automotive spare parts for two years in monthly buckets. The empirical distributions of non-zero demand and inter-demand

[2]It is possible to produce a constant forecast variant, closer to the original Croston's method. However, no advantages for this were identified and only the dynamic version is discussed here.

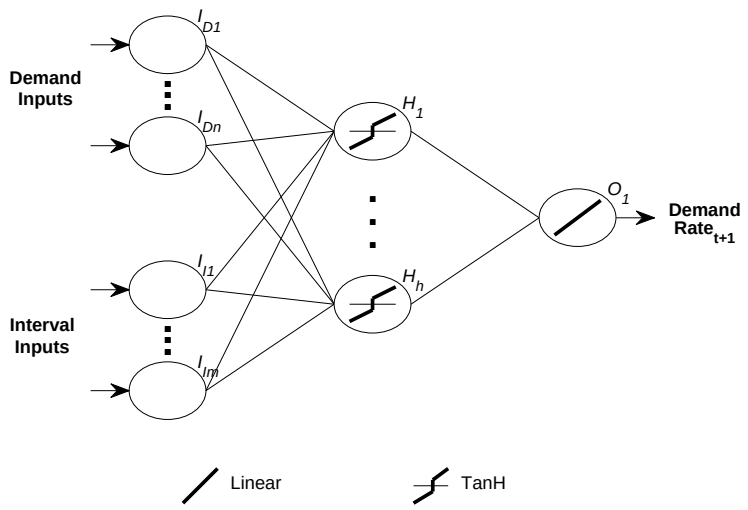

Figure 2: The $NN - Rate$ architecture, with variable number of demand I_{Dn} and interval I_{Im} lagged inputs and hidden nodes H_h, which use the TanH activation function. A single linear output provides the demand rate forecast.

intervals were estimated.[3] Based on those, new monthly intermittent time series were constructed. Each time series has 236 observations, out of which 36 observations, 3 years of history, are used as in-sample data, 100 are used as out-of-sample data over which the performance of the different models is evaluated and the remaining 100 are used as burn-in period for the simulation. Since the models are assessed not only in terms of forecasting accuracy but also using inventory metrics, it is necessary to simulate the stocks and orders for each item. In order to achieve realistic levels for each, irrespective of initial stock and orders, the simulation is run for the burn-in period before any statistics are collected from the out-of-sample evaluation period. This allows each model to reach its normal behaviour and stock levels.

It was argued in section 1 that the independence between the non-zero demand and the inter-demand interval is often taken for granted, which may not always hold. Section 2 presented NN models capable of capturing dependence between the variables in a linear or nonlinear way. Figure 3 provides

[3]Well known distributions provided a poor fit.

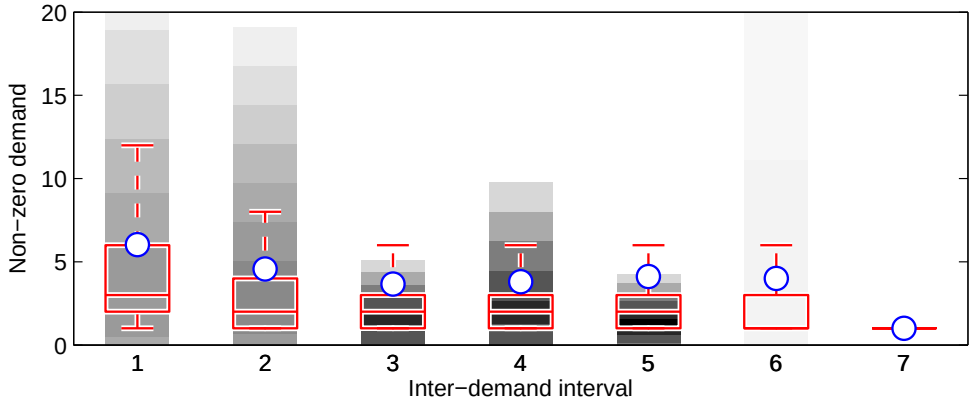

Figure 3: Boxplot of non-zero demands for each inter-demand interval. Mean of the distribution is plotted as a circle, while background colouring is the density as estimated by a Gaussian kernel. Darker areas signify higher density.

boxplots of the non-zero demand for each inter-demand interval of the original dataset from Syntetos and Boylan (2005). The data for the boxplots has been pooled across all time series of the dataset. The mean of each group of non-zero demand has been plotted as a circle. The density of the distribution corresponding to each boxplot has been estimated using Gaussian kernel density estimation and is provided as grayed background for each boxplot. Darker areas signify higher density. The plot has been capped to demands up to 20, as higher demands up to 416 are outliers and would make reading it harder. It is apparent that the distribution of non-zero demand changes for different inter-demand intervals; hence they are not independent as assumed by Croston's method. Furthermore, the relationship between z_t and x_t is not linear. A 2^{nd} order polynomial was identified to fit better the data, specified automatically using the Bayesian Information Criterion, revealing the presence of significant nonlinearities. Therefore, the NNs are expected to be able to capture this nonlinear relationship, thus improving their performance. Note that NNs, as universal approximators, are not restricted by the nature of the nonlinearity.

3.2. Accuracy Metrics

Measuring forecasting accuracy for intermittent demand time series is not straightforward. Although it is trivial to measure the Mean Absolute Error

(MAE) for each time series separately, in order to summarise the results across several time series scale-independent errors must be used. Intermittent time series have zero values making relative-to-the-series scale independent errors, such as the Mean Absolute Percentage Error (MAPE), impossible to calculate. Relative errors, such as the Geometric Mean Relative Absolute Error (GMRAE), require the calculation of a benchmark model, which is often the Naive. This can result in zero denominator, making the calculation of such errors impossible. Syntetos and Boylan (2005) argue that an error measure that is robust to such problems is the Relative Geometric Root Mean Square Error (RGRMSE) based on the work by Fildes (1992). However, for this set of experiments, due to the naive based benchmarks and the time series models, it is possible to have zero individual errors, which consequently in zero geometric mean error. Hyndman and Koehler (2006) proposed the use of Mean Absolute Scaled Error (MASE) instead, which is essentially the out-of-sample MAE of the method to be evaluated, across the relevant forecast horizon, divided by the in-sample one-step ahead Naive forecast. This error measure minimises the probability of zero or infinite error due to calculation. However, its interpretation is very unclear, as well as what the relative differences between different sizes of the error mean. Therefore a fully satisfactory error measure for intermittent demand is not readily available.

If the difference in magnitude between the error of different methods is not required and the focus is on the ranking of the models then using average ranks of MAE has several advantages. MAE is calculated as $\sum |e_t|/n$, where $e_t = A_t - F_t$ and n the number of samples; A_t being the actual demand for period t and F_t the forecast for the same period. MAE is not biased in any way and does not suffer from any problems with zero values or with aggregation of individual errors that may be zero. Using ranks of MAE overcomes its scale dependent nature and it is possible to summarise results across different time series. The ranking of each forecasting model for each time series is calculated and then the average ranks across all time series. This becomes a powerful non-parametric accuracy error, being easily incorporated in the calculation of non-parametric statistical tests that can highlight significant differences between the models. In this study more than two models are compared and therefore pairwise tests cannot be used, as the multiple comparisons will make statistical inference misleading. Demšar (2006) discusses alternative methods to compare several models simultaneously. Demšar suggests first using the Friedman test, a non-parametric analogue to ANOVA that identifies if at least one of the models is different from the rest. If signif-

icant differences are identified, then one should apply the post-hoc Nemenyi test to rank the models and identify further differences. A similar approach is used to aggregate Mean Error (ME) figures across time series to measure forecasting bias. ME is calculated simply as $\sum e_t/n$.

All forecasts are three-steps ahead, as it will be discussed in detail in the inventory simulation design. Therefore 98 out-of-sample multi-step ahead forecasts are produced and their respective errors calculated using MAE, in a rolling origin scheme as discussed by Tashman (2000).

3.3. Inventory Simulation Design

In order to calculate inventory metrics, an inventory simulation is setup. The commonly applied in practice order-up-to policy (T,S) is used, as per Teunter and Sani (2009). The review time is set to monthly ($T = 1$), while the order-up-to level S can be calculated as:

$$S = \hat{D} + k\hat{\sigma_L}, \tag{11}$$

where $\hat{D}$ is the demand over the lead time period, k is a safety factor for achieving target service level and $\hat{\sigma_L}$ is the variance of the error of the forecasts for the respective lead time L. $\hat{D}$ is conventionally calculated as LY'_{t+1}, i.e. the one-step ahead demand forecast is multiplied by the lead time to find the total demand over the period. This calculation is fine as long as the demand forecast is constant, which is true for Croston's method and its modifications. However this is not true for the proposed neural network models that can output dynamic forecasts, whose values are not equal over different forecast horizons. Therefore $\hat{D} = \sum_{i=1}^{L} Y'_{t+i}$, i.e. the sum of the forecasted demand over a given forecast horizon (or supply lead time). $\hat{\sigma_L}$ for multi-step forecasts is approximated as $\sqrt{L}\hat{\sigma_1}$, where $\hat{\sigma_1}$ is the one-step ahead forecast standard error, i.e. the square root of the mean squared error (MSE). The approximation can be avoided by directly calculating the empirical MSE for the relevant forecast horizon. Finally, k is calculated from the normal distribution, depending on the desired service level. In this simulation the lead time is set $L = 3$, forcing the forecast horizon to be three periods as well. Service levels for 0.80, 0.90, 095 and 0.99 are considered.

In each period of the simulation, the realised demand for each item is subtracted from the holding stock H. If the stock falls below S, then an order $S - H$ is placed with a lead time $L = 3$. If the order cannot be serviced, an out-of-stock event is measured. To simplify the simulation, unserviced orders

are considered lost. Service level α and β are measured. The former measures the probability of not having a stock-out event, while the latter measures the magnitude in units of serviced demand over the full expected demand. By definition service level $\alpha \leq$ service level β. Furthermore, holding volume and backlog volumes are tracked. These metrics allow us to consider trade-off curves between stock-holding and backlog volumes and holding stock and service levels as in Teunter et al. (2010) and Syntetos et al. (2010). Both holding and backlog volumes are scaled by the average non-zero demand for each time series. This way they become scale-independent and therefore it is possible to summarise them across time series.

To initialise the simulation each item is assumed to have full stock, $H = S$. However, to mitigate any bias from this assumption the simulation has a "'burn-in"' period of 100 iterations before any statistics are collected. This brings the initial holding stocks of the out-of-sample period to reasonable levels for each forecasting model. The results that follow in the next section are based on data after this initial "'burn-in"' period.

3.4. Methods

Forecast are created for each item based on the methods discussed in section 2. All Croston's variants based on exponential smoothing are run modelled with $\alpha = \{0.05, 0.10, 0.15, 0.20\}$. Identical values for the smoothing parameter are used for SES. All MA and moving average Croston's variants are modelled using $k = \{3, 5\}$.

For the NNs three different settings for the number of input lags and hidden nodes are used. Both I and H take values from 1 to 3. These are retained to relatively small values in order to keep the model degrees of freedom into a reasonable range in comparison to the size of the in-sample data. This way, the maximum degrees of freedom for three lags for each input and three hidden nodes is 22. All other network settings are kept constant. Networks are regularised and a $\gamma = 0.9$ is used. The networks are trained using the Levenberg-Marquardt algorithm, which requires setting the μ_{LM} and its increase and decrease steps. Here $\mu_{LM} = 10^{-3}$, with an increase step of $\mu_{inc} = 10$ and a decrease step of $\mu_{dec} = 10^{-1}$. For a detailed description of the algorithm and the parameters see Hagan et al. (1996). This training algorithm allows for fast training, essential for large scale forecasting applications. The maximum training epochs are set to 1000. The training can stop earlier if μ_{LM} becomes equal or greater than $\mu_{max} = 10^{10}$. The training of all networks is initialised with random weights 5 times to avoid getting trapped

in a bad local minimum. Note that a large number of initialisations is not needed due to the regularisation, that forces the weights to converge to small values. Since no validation sample is required for training, it is not possible to select between the different training initialisations without overfitting to the training set. To avoid this an ensemble over all initialisations is considered. The median of the forecasts of all five training initialisations is the final output. This step robustifies further the models against overfitting, which is crucial for small samples. Finally, all inputs are linearly scaled between $[-0.8, 0.8]$.

The network model proposed by Gutierrez et al. (2008) is also used as a benchmark, named $NN - GSM$, with one significant modification. Instead of using the standard back-propagation with momentun, regularised loss and the Levenberg-Marquardt training algorithm are used to allow training the networks in small samples, similar to the other NNs. Furthermore, an ensemble of 5 training initialisations is used to produce forecasts, similar to the other network models. These changes substantially increases the robustness of the model. Initial results using the original training method were very poor and were discarded from the study. The poor results were attributed to the small training sample.

Including all the parameter combinations, in total 45 models are simulated in this study. To facilitate the analysis of the results, only the best performing parameters of each model will be presented in section 4. The criteria for selecting the model is minimum in-sample mean MAE rank, as realistically it is impossible to pick a model on out-of-sample statistics, including inventory metrics.

4. Results

4.1. Accuracy Metrics

Table 1 provides the average ranks for in- and out-of-sample ME and MAE. In brackets the relative model rank is provided. Friedman test for both ME and MAE indicated significant differences between the models with p-value of 0.000. The critical distance for the Nemenyi test was found 0.66, 0.58 and 0.54 for significance levels 1%, 5% and 10% respectively. Models with average ranks different more than the critical distance are significantly different in performance. Figure 4 provide visually the results of the Nemenyi test at 5% significance. Models grouped by the vertical brackets have no evidence of statistically significant differences.

Considering the ME, a striking result is that the conventional time series models, namely the $Naive$, MA and SES, are the least biased. Croston's method and its modifications follow. The severely biased $CR-SES$ is corrected by both $CR-SES-SB$ and $CR-SES-SH$ in-sample. However, out-of-sample this is not true. The $CR-MA$ method performs well, while its corrections do not improve its bias. All neural networks are strongly biased. Although table 1 gives the impression of different rankings between the models, consulting the Nemenyi results in figure 4 reveals that there is no evidence to support significant differences in many cases, in particular for the NNs.

Table 1: Accuracy Metrics

Model	rank of ME		rank of MAE	
	in-sample	out-of-sample	in-sample	out-of-sample
Naive	5.62 (1)	2.60 (1)	20.10 (13)	18.65 (13)
MA	6.67 (2)	2.98 (2)	15.27 (11)	12.02 (9)
SES	7.56 (3)	3.25 (3)	13.12 (9)	9.13 (7)
CR-Naive	9.31 (6)	5.11 (4)	18.44 (12)	16.42 (11)
CR-MA	7.97 (4)	6.04 (6)	15.10 (10)	11.63 (8)
CR-MA-SB	10.32 (8)	11.84 (9)	10.18 (6)	6.76 (5)
CR-MA-SH	12.58 (10)	13.56 (10)	8.61 (2)	4.93 (3)
CR-SES	15.83 (13)	5.26 (5)	10.42 (7)	4.96 (4)
CR-SES-SB	10.05 (7)	7.49 (7)	9.52 (4)	4.57 (2)
CR-SES-SH	9.26 (5)	9.72 (8)	9.07 (3)	4.53 (1)
NN-GSM	12.33 (9)	14.93 (12)	9.99 (5)	14.88 (10)
NN-Dual	14.21 (12)	13.81 (11)	5.32 (1)	9.08 (6)
NN-Rate	12.74 (11)	19.14 (13)	12.83 (8)	18.21 (12)

Considering the MAE a different picture emerges. In-sample the $NN-Dual$ perform best. The Croston's method variants follow with the conventional time series models averaging last, as expected. Out-of-sample the ranking differs substantially. The modified Croston, both for $CR-SES$ and $CR-MA$ rank first. However, figure 4 shows that only $CR-SES-SH$ is statistically different from the original method $CR-SES$. SES follows together with $NN-Dual$ after almost all Croston's modifications. The remaining NN models follow. Similar to the result for ME, $NN-Rate$ ranks almost last, outperforming only the naive.

Based on the results from the accuracy metrics there is little benefit

of NN for intermittent demand over the original and modified Croston's method. The same is true for the $NN-GSM$ model proposed by Gutierrez et al. (2008), in contrast to their findings. However, it it important to note that this study uses short monthly intermittent time series instead of daily long lumpy demand series, as in their original study, that may explain the different findings.

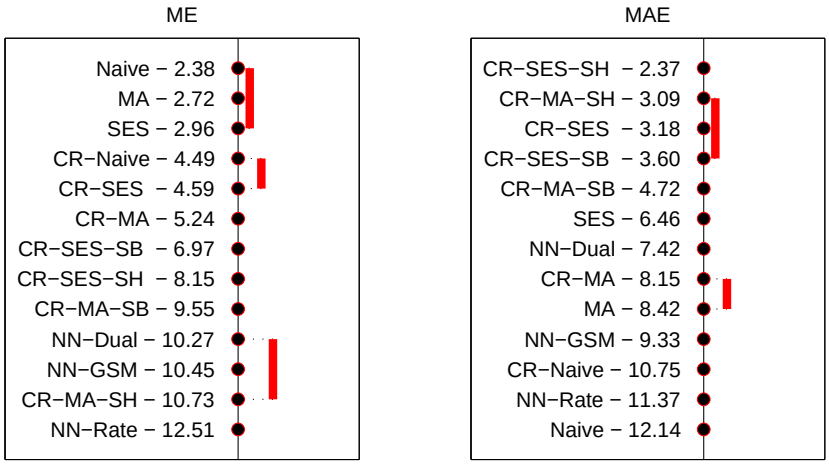

Figure 4: Nemenyi test results for out-of-sample ranks of ME and MAE. For model included in vertical brackets there is no evidence of significant differences at 5% level.

4.2. Inventory Metrics

The story presented by inventory metrics is substantially different to the accuracy metrics. Service levels α and β as well as scaled holding and backlog volumes at the end of the simulation are provided for four different target service levels in tables 2 and 3 respectively. Models can have high service levels by retaining more stock, therefore evaluating solely the service levels separately does not provide any insight. To overcome this the inventory-backlog efficiency is examined. This trade-off curve between scaled holding and backlog volume shows at what cost of unmet demand the holding volume is kept low.

Examining the time series models it is apparent that the high service level of the $Naive$ and $CR-Naive$ comes at a substantial cost in stock holding. MA and SES perform very similarly considering service levels and

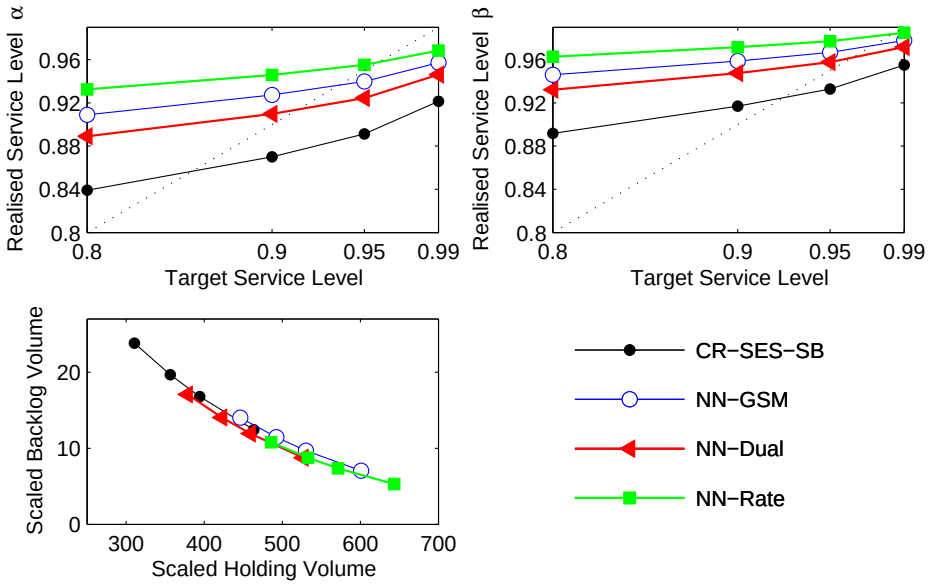

Figure 5: Realised service levels α and β against target service and scaled backlog volume and scaled holding volume trade-off curve for best performing models.

Table 2: Realised Service Levels α (β).

Model	Target Service Level			
	80%	90%	95%	99%
Naive	95.0% (97.1%)	96.2% (97.9%)	97.0% (98.5%)	98.1% (99.1%)
MA	87.4% (91.8%)	90.0% (93.7%)	91.7% (95.0%)	94.0% (96.7%)
SES	86.6% (91.2%)	89.2% (93.2%)	90.9% (94.5%)	93.5% (96.3%)
CR-Naive	92.6% (95.5%)	94.3% (96.7%)	95.4% (97.5%)	97.0% (98.4%)
CR-MA	86.4% (91.0%)	89.1% (93.2%)	91.0% (94.5%)	93.6% (96.4%)
CR-MA-SB	82.8% (88.2%)	86.2% (91.0%)	88.5% (92.8%)	91.8% (95.2%)
CR-MA-SH	82.1% (87.6%)	85.6% (90.5%)	88.0% (92.4%)	91.4% (95.0%)
CR-SES	84.6% (90.0%)	87.5% (92.3%)	89.5% (93.7%)	92.4% (95.8%)
CR-SES-SB	83.9% (89.2%)	87.0% (91.7%)	89.1% (93.3%)	92.2% (95.5%)
CR-SES-SH	83.7% (89.0%)	86.8% (91.6%)	89.0% (93.2%)	92.0% (95.4%)
NN-GSM	90.9% (94.6%)	92.7% (95.9%)	94.0% (96.7%)	95.7% (97.8%)
NN-Dual	88.9% (93.2%)	91.0% (94.8%)	92.4% (95.8%)	94.6% (97.2%)
NN-Rate	93.2% (96.3%)	94.6% (97.2%)	95.5% (97.7%)	96.8% (98.5%)

Table 3: Scaled Holding (Backlog) Volume.

Model	Target Service Level			
	80%	90%	95%	99%
Naive	693.0 (8.1)	750.5 (6.2)	797.6 (5.0)	885.2 (3.3)
MA	396.8 (18.8)	446.0 (15.4)	485.8 (13.0)	559.2 (9.5)
SES	361.6 (20.1)	408.4 (16.6)	446.5 (14.2)	516.7 (10.4)
CR-Naive	617.8 (11.3)	672.3 (8.9)	716.7 (7.3)	798.6 (5.0)
CR-MA	374.5 (20.1)	422.8 (16.5)	462.0 (14.0)	534.3 (10.2)
CR-MA-SB	309.3 (24.9)	357.1 (20.5)	396.1 (17.4)	467.8 (12.8)
CR-MA-SH	296.7 (25.9)	344.5 (21.4)	383.5 (18.2)	455.4 (13.4)
CR-SES	291.3 (23.3)	335.2 (19.3)	371.1 (16.5)	437.6 (12.2)
CR-SES-SB	310.6 (23.8)	356.7 (19.7)	394.2 (16.8)	463.6 (12.4)
CR-SES-SH	306.9 (24.1)	353.0 (19.9)	390.5 (17.0)	459.9 (12.6)
NN-GSM	446.2 (14.0)	492.3 (11.5)	530.3 (9.7)	601.1 (7.1)
NN-Dual	378.5 (17.1)	422.8 (14.1)	459.1 (11.9)	526.9 (8.8)
NN-Rate	485.7 (10.8)	532.8 (8.8)	571.4 (7.4)	643.3 (5.3)

stock efficiency. $CR-SES$ and its modifications appear to achieve higher service levels for the same stock holding level in comparison to $CR-MA$ and its modifications, however there are no substantial differences between them. The $CR-SES-SB$ is chosen as the best model of this family.

The comparison between the best statistical model with the NNs is facilitated in figure 5. Figure 5 plots the service levels and trade-off curve between the best models of all different model groups in order to avoid cluttering the graphs. Considering the realised service levels α and β, curves that are higher dominate others. The dotted diagonal line is the ideal performance, where the realised and the target service levels match. The proposed $NN-Rate$ achieve the highest realised service level for all different target service levels, followed by $NN-GSM$ and $NN-Dual$. NNs overall dominate Croston based methods. The third subplot in figure 5 provides the trade-off curves between scaled holding and backlog volumes. Curves that are closer to the bottom-left corner of the plot dominate others, as they achieve lower backlogs with lower holding volume. Here all four models appear to be on the same level. Therefore, $NN-Rate$ offers substantially better realised service levels without needing to jump to a different trade-off holding and backlog curve. Notably, $NN-Dual$ trade-off curve marginally dominates other models, while providing higher service levels than the best Croston's variant.

4.3. Discussion

The results suggest that NNs are good contesters for intermittent demand forecasting problems. This work proposed a framework to construct robust networks for short intermittent time series that are prevalent in practice, extending on the work of Gutierrez et al. (2008) who demonstrated good performance of NNs in the presence of long daily lumpy time series. The proposed NNs are able to capture the interaction between the non-zero demand and the inter-demand intervals that is assumed non-existent by Croston's method and its variants. Furthermore the proposed models are able to capture the dynamics of the time series and therefore are not restricted to outputting constant forecasts. Although $NN-Dual$ still requires de-biasing of the forecasts, $NN-Rate$ do not, as no inversion bias in introduced. $NN-Rate$ was found to be the best performing model, reaching substantially higher service levels with minimal small increase in the holding volumes, while its corresponding inventory trade-off curve was at the same level of other competing models.

The simulation in this study used both accuracy and inventory metrics. Inventory metrics, such as service levels and holding stock, are directly related to organisations' inventory performance and are closely related to decision making. On the other hand, forecasting accuracy metrics are more abstract. However, these are often assumed to be related to inventory performance (Tashman, 2000) and often the forecasting literature stops at reporting only accuracy metrics. Levén and Segerstedt (2004) showed that keeping track of both accuracy and inventory metrics can lead to insightful findings for intermittent demand. This paper provided further evidence of the usefulness of investigating both accuracy and inventory metrics. In particular, findings suggest that accuracy metrics alone, can sometimes lead to misleading conclusions for intermittent demand. Inventory metrics were evaluated following the suggestions of the literature(for e.g. see Levén and Segerstedt, 2004; Eaves and Kingsman, 2004; Syntetos and Boylan, 2006). For instance, Eaves and Kingsman (2004) reach a similar conclusion. Selection of a forecasting method by forecasting accuracy and inventory metrics differ. Forecasting accuracy should be distinguished from the stock control performance of a model. There is a substantial body of literature discussing this issue (for e.g. see Gardner, 1990; Sani and Kingsman, 1997; Syntetos and Boylan, 2006; Strijbosch et al., 2011), raising the importance of evaluating inventory metrics, beyond forecasting accuracy. The disconnect between forecasting accuracy and inventory metrics is also illustrated in the argument by (Stri-

jbosch et al., 2011); it is common practice to optimise forecasting models on the one-step ahead forecasting error, while the models are deployed to aid longer lead time inventory decisions. The disconnect between the two is discussed thoroughly in the previous references.

In this study, based on robust accuracy metrics NN showed poor forecasting performance, while the opposite was found true, based on the inventory simulation. $NN - Rate$ outperformed other alternatives. Syntetos and Boylan (2005) and Hyndman and Koehler (2006) argued that measuring forecasting accuracy for intermittent demand time series is not trivial. One has to consider the presence of zeros in the demand, and potentially in the forecasting errors. This can make many well established error measures impossible to calculate and weaken the reliability of those that can be calculated. For instance, Teunter and Duncan (2009) concluded that mean absolute deviation favoured under-forecasting. Furthermore, measuring forecasting performance for intermittent demand time series in the conventional way can be shown to erroneous in a more fundamental way and should not be used for model selection purposes. Croston's method and its variants, as well as the NNs proposed here, do not output the point forecast demand for each period, rather a "'demand-rate"' as discussed in section 2. Consequently, measuring the deviation of the raw time series against the demand rate forecast of the models is not meaningful. To illustrate this point further, consider the following example; suppose that for a time series the demand for the next three periods is (0,0,9). The optimum Croston's forecast would be a demand rate per period of three units, or (3,3,3) for the next three periods, resulting in zero excess stock and covering fully the demand. Such a forecast can be interpreted as a demand of 9 units distributed over a number of periods, since the exact timing of the demand event is unknown. However, no matter which error metric is employed there will always be forecasting error. This is due to the focus of conventional metrics on individual time periods and the importance they put on the timing of the demand. Wallström and Segerstedt (2010) propose two novel measures of bias, the *number of shortages* and *periods in stock* that avoid this issue. The advantage of these measures is that they are based on notions of cumulative error, therefore not focusing on a specific time periods, as conventional error measures do. Therefore, more research in such kind of error metrics and their implementation in practice is desirable.

5. Conclusions

This study proposes a series of neural network models for intermittent demand forecasting. Two models are proposed, namely the $NN-Dual$ and $NN-Rate$. Both are bivariate models that allow interactions between the demand and the inter-demand intervals of intermittent items, an extension to conventional intermittent demand modelling. The networks employ regularised cost functions and median ensembles to produce robust forecasts and make them applicable to time series with short in-sample history that is often the case for intermittent demand data. The performance of the models is measured using both accuracy and inventory metrics. The results are conflicting and it is argued that the accuracy metrics provide misleading findings and in fact are erroneous for intermittent demand data. Based on realised service levels and trade-off curves between holding and backlog volumes the proposed neural network models out-perform conventional Croston's method and its modifications. $NN-Rate$ achieves the highest service rates in the simulation at the expense of small increase in the holding stock, which is however much lower than time series methods that reach similarly high service levels by overstocking, such as the $Naive$ and $CR-Naive$.

An inventory simulation of 1000 time series, based on the dataset by Syntetos and Boylan (2005) is used to provide empirical evidence of the neural networks' performance. Further simulations, particularly with different levels of intermittency, will allow exploring further the conditions under which the proposed neural networks perform well for intermittent time series and explore further the effects of training sample size for their performance.

References

Adya, M., Collopy, F., 1998. How effective are neural networks at forecasting and prediction? A review and evaluation. Journal of Forecasting 17 (5-6), 481–495.

Bishop, C. M., 1996. Neural Networks for Pattern Recognition, 1st Edition. Oxford University Press, USA.

Boylan, J., Syntetos, A., 2007. The accuracy of a modified croston procedure. International Journal of Production Economics 107 (2), 511 – 517.

Connor, J. T., Martin, R. D., Atlas, L. E., 1994. Recurrent neural networks and robust time series prediction. IEEE Transactions on Neural Networks 5, 240–254.

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

Dahl, C. M., Hylleberg, S., 2004. Flexible regression models and relative forecast performance. International Journal of Forecasting 20 (2), 201–217.

Demšar, J., December 2006. Statistical comparisons of classifiers over multiple data sets. The Journal of Machine Learning Research 7, 1–30.

Eaves, A. H. C., Kingsman, B. G., 2004. Forecasting for the ordering and stock-holding of spare parts. The Journal of the Operational Research Society 55 (4), 431–437.

Fildes, R., June 1992. The evaluation of extrapolative forecasting methods. International Journal of Forecasting 8 (1), 81–98.

Gardner, Jr., E. S., April 1990. Evaluating forecast performance in an inventory control system. Management Science 36, 490–499.

Ghobbar, A. A., Friend, C. H., December 2003. Evaluation of forecasting methods for intermittent parts demand in the field of aviation: a predictive model. Comput. Oper. Res. 30, 2097–2114.

Gutierrez, R. S., Solis, A. O., Mukhopadhyay, S., 2008. Lumpy demand forecasting using neural networks. International Journal of Production Economics 111 (2), 409 – 420.

Hagan, M. T., Demuth, H. B., Beale, M. H., 1996. Neural Network Design. MA: PWS Publishing, Boston.

Hornik, K., 1991. Approximation capabilities of multilayer feedforward networks. Neural Networks 4 (2), 251 – 257.

Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Networks 2 (5), 359 – 366.

Hyndman, R. J., Koehler, A. B., 2006. Another look at measures of forecast accuracy. International Journal of Forecasting 22 (4), 679–688.

Hyndman, R. J., Shenstone, L., 2005. Stochastic models underlying Croston's method for intermittent demand forecasting. Journal of Forecasting 24 (6), 389–402.

Johnston, F. R., Boylan, J. E., 1996. Forecasting for items with intermittent demand. The Journal of the Operational Research Society 47 (1), pp. 113–121.

Johnston, F. R., Boylan, J. E., Shale, E. A., 2003. An examination of the size of orders from customers, their characterisation and the implications for inventory control of slow moving items. The Journal of the Operational Research Society 54 (8), pp. 833–837.

Levén, E., Segerstedt, A., 2004. Inventory control with a modified croston procedure and erlang distribution. International Journal of Production Economics 90 (3), 361 – 367.

MacKay, D. J. C., 1992. Bayesian interpolation. Neural computation 4, 415–447.

Markham, I. S., Rakes, T. R., April 1998. The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. Computers & Operations Research 25, 251–263.

Rao, A., 1973. A comment on: Forecasting and stock control for intermittent demands. Operational Research Quarterly 24, 639–640.

Rumelhart, D., Hinton, G., Williams, R., 1988. Parallel Distributed Processing Explorations in the Microstructure of Cognition. MIT Press.

Sani, B., Kingsman, B. G., 1997. Selecting the best periodic inventory control and demand forecasting methods for low demand items. Journal of the Operational Research Society 48 (7), 700–713.

Shale, E. A., Boylan, J. E., Johnston, F. R., 2006. Forecasting for intermittent demand: the estimation of an unbiased average. The Journal of the Operational Research Society 57, 588–592.

Strijbosch, L. W. G., Syntetos, A. A., Boylan, J. E., Janssen, E., 2011. On the interaction between forecasting and stock control: The case of non-stationary demand. International Journal of Production Economics 133 (1), 470–480.

Syntetos, A. A., Boylan, J. E., May 2001. On the bias of intermittent demand estimates. International Journal of Production Economics 71 (1-3), 457–466.

Syntetos, A. A., Boylan, J. E., 2005. The accuracy of intermittent demand estimates. International Journal of Forecasting 21 (2), 303 – 314.

Syntetos, A. A., Boylan, J. E., 2006. On the stock control performance of intermittent demand estimators. International Journal of Production Economics 103 (1), 36 – 47.

Syntetos, A. A., Boylan, J. E., Croston, J. D., 2005. On the categorization of demand patterns. Journal of the Operational Research Society 56 (5), 495–503.

Syntetos, A. A., Nikolopoulos, K., Boylan, J. E., 2010. Judging the judges through accuracy-implication metrics: The case of inventory forecasting. International Journal of Forecasting 26 (1), 134 – 143.

Tashman, L. J., 2000. Out-of-sample tests of forecasting accuracy: an analysis and review. International Journal of Forecasting 16 (4), 437 – 450.

Teunter, R., Sani, B., 2009. Calculating order-up-to levels for products with intermittent demand. International Journal of Production Economics 118 (1), 82 – 86.

Teunter, R., Syntetos, A., Babai, M., June 2010. Determining order-up-to levels under periodic review for compound binomial (intermittent) demand. European Journal of Operational Research 203 (3), 619–624.

Teunter, R. H., Duncan, L., 2009. Forecasting intermittent demand: a comparative study. The Journal of the Operational Research Society 60, 321–329.

Wallström, P., Segerstedt, A., 2010. Evaluation of forecasting error measurements and techniques for intermittent demand. International Journal of Production Economics 128 (2), 625–636.

Willemain, T. R., Smart, C. N., Schwarz, H. F., 2004. A new approach to forecasting intermittent demand for service parts inventories. International Journal of Forecasting 20 (3), 375 – 387.

Willemain, T. R., Smart, C. N., Shockor, J. H., DeSautels, P. A., 1994. Forecasting intermittent demand in manufacturing: a comparative evaluation of Croston's method. International Journal of Forecasting 10 (4), 529–538.

Zhang, G., Patuwo, B. E., Hu, M. Y., 1998. Forecasting with artificial neural networks: The state of the art. International Journal of Forecasting 14 (1), 35 – 62.