Some forecasting principles from the M4 competition

Jennifer L. Castle, Jurgen A. Doornik, and David F. Hendry*
Magdelen College and Nuffield College

January 9, 2019

Abstract

Economic forecasting is difficult, largely because of the many sources of nonstationarity. The M4 competition aims to improve the practice of economic forecasting by providing a large data set on which the efficacy of forecasting methods can be evaluated. We consider the general principles that seem to be the foundation for successful forecasting, and show how these are relevant for methods that do well in M4. We establish some general properties of the M4 data set, which we use to improve the basic benchmark methods, as well as the Card method that we created for our submission to the M4 competition. A data generation process is proposed that captures the salient features of the annual data in M4.

Automatic forecasting, Calibration, Prediction intervals, Regression, M4, Seasonality, Software, Time series, Unit roots

1 Introduction

Economic forecasting is challenging. No clear consensus approach has arisen in the literature. Sophisticated methods often fail to beat a simple autoregressive model. Then, when a small advantage is shown, this may fail to survive in slightly different settings or time periods. What does hold is that all theoretical results that assume stationarity are irrelevant. Instead, small shocks occur regularly, and large shocks and structural breaks intermittently. Causes of such breaks could be financial crises, trade wars, conflicts, policy changes, etc. So nonstationarities can arise from unit roots as well as structural breaks in mean or variance.

The M3 and M4 competitions create realistic, immutable, and shared data sets that can be used as testbeds for forecasting methods. They have made invaluable contributions to improving the quality of economic forecasting, as well as increasing our understanding of methods and techniques. There are limitations to the data: variables are anonymized, and have unknown and differing sample periods. This prevents the use of subject matter expertise (or even judgement). It also creates problems for methods that try to link variables for forecasting: the misaligned sample periods may cause some variables to be ahead in time of others. Unfortunately, this also rules out any multiple variable or factor approaches. A useful addition to M4 is the request for forecast intervals: good expression of the uncertainty of a forecast is just as important as the forecast itself. Probabilities of rain are now a routine aspect of weather forecasts, and forecast uncertainty should be quantified in other settings as well.

Over time, between our research, the literature, and M4 results, there seem to emerge a few relatively general 'principles' for economic forecasting:

*Financial support from the Robertson Foundation (award 9907422) and Institute for New Economic Thinking (grant 20029822) is gratefully acknowledged.

A] dampen trends/growth rates;
B] average across forecasts from 'non-poisonous' methods;
C] include forecasts from robust devices in that average;
D] select variables in forecasting models at a loose significance;
E] update estimates as data arrive, especially after forecast failure.

The authors made a submission to the M4 competition that did well in many aspects. Here we aim to interpret our results, as well as some methods that did well in M3 and M4, in the light of these principles. As part of this we summarize the properties of the M4 data set, which leads to improved benchmark forecast methods.

Our proposed forecast method, *Card*, is formally described in Doornik, Castle, and Hendry (2019). The procedure is based on simple autoregressive models, augmented with a dampened trend and some robustification through differencing. Forecast combination is also used. Below we show that some further improvement can be made by paying more attention to our principles. Furthermore, improved formulation of the forecast standard errors is provided.

M4 is the fourth generation of M forecast competitions, created by Spyros Makridakis and the M4 team. M4 provides a database of $100\,000$ series requiring out-of-sample forecasts. This large size makes it a computational challenge too. Efficient production of forecasts is useful, even more so when studying subsample properties. The previous competition, M3, consisted of 'only' 3003 variables. The best performing method in the M3 competition is the so-called *Theta* method, see Makridakis and Hibon (2000) and §2.1 below, so many papers take that as the benchmark to beat.

Proper handling of seasonality is expected to be an important aspect of forecasting. This implies that the exercise has some similarities to the X12-ARIMA programme of the US Census Bureau, see Findley, Monsell, Bell, Otto, and Chen (1998). The X12 approach involves estimating a seasonal ARIMA model to extend the series with forecasts, followed by smoothing using a sequence of moving averages in the seasonal and deseasonalized directions. Moving averages are sensitive to outliers and structural breaks, and procedures need to make allowance for this.

The remainder of this paper is as follows. We discuss the stucture of M4 and its benchmark methods in §2, together with data visualization in §3. Next, we adapt the Theta method, and also introduce a simple but effective variant in §4. We also consider the expected outcomes of the accuracy measures, and study the performance of the new benchmark methods. In §5 we consider improvements to the *Card* method. In §6 we propose a simulation experiment that captures the salient features of yearly M4. Finally, §7 concludes. Derivations are presented in appendices.

2 M4 competition

Some basic aspects of the M4 data are given in Table 1, including values of the forecast horizon H and frequency S. The data is also classified into six categories: demographic, finance, industry, macro, micro, and other. This information is not used in our methods. The yearly, monthly, and quarterly series together constitute 95% of the sample, so will dominate the overall results.

The frequency is based on the labels 'hourly', 'weekly' etc. but is not otherwise provided (so daily data could be for five weekdays or a full seven day week). The frequency m is used[1] in the performance measures (§2.3), but S in estimation (with some exceptions). A second frequency S_2 can capture longer cycles. For hourly data, $S = 24$ reflects the diurnal rhythm, while $S_2 = 7$ creates an additional frequency of $SS_2 = 168$ for the weekly pattern. If there is enough data, an annual pattern could be added: energy consumption, e.g., is quite different during public holidays.

[1] We only discovered the value of m after the end of the competition.

	dimension		frequency		sample size		forecasts	T_{limit}
	# series	%	$S \times S_2$	m	T_{min}	T_{max}	H	
Yearly	23000	23.0	1	1	13	835	6	40 years
Quarterly	24000	24.0	4	4	16	866	8	40 years
Monthly	48000	48.0	12	12	42	2794	18	40 years
Weekly	359	0.4	52	1	80	2597	13	40 years
Daily	4227	4.2	1	1	93	9919	14	3650
Hourly	414	0.4	24×7	24	700	960	48	5040

Table 1: Basic properties of the M4 data set

The sample sizes range from 13 annual observations to almost ten thousand daily observations. Table 1 records the lengths of shortest and longest series as T_{min} and T_{max}. The sample sizes refer to the competition (or training) version. The objective is to create H forecasts beyond the end of the sample. The M4 organizers held back the outcomes in order to evaluate the submitted forecasts. These were subsequently made available as a separate data set. When developing our methods, we emulated the evaluation procedure by withholding a further H observations.

2.1 M4 benchmark forecasting methods

Several forecasting techniques are used in M4 as benchmark methods. We review random walk, exponential smoothing, and *Theta* forecasts, but ignore the basic machine learning and neural network approaches.[2]

The random walk forecasts of annual and nonseasonal data are a simple extrapolation of the last observation. This is called *Naive2* forecasts in M4:

$$\widehat{y}_{T+h} = y_T, h = 1, ..., H.$$

Exponential smoothing (ES) methods are implemented as single source of innovation models, see Hyndman, Koehler, Snyder, and Grose (2002) and Hyndman, Koehler, Ord, and Snyder (2008). They can be formulated using an additive or multiplicative (or mixed) representation. The additive exponential smoothing (AES) model has the following recursive structure:

$$\begin{aligned}
\mu_t &= l_{t-1} + b_{t-1}, \\
\epsilon_t &= y_t - \mu_t, \\
l_t &= l_{t-1} + \alpha\epsilon_t, \\
b_t &= b_{t-1} + \delta\epsilon_t,
\end{aligned}$$

where y_t is the observed time series, ϵ_t the one-step prediction error, l_t the level, and b_t the slope. Given initial conditions l_0 and b_0, the coefficients α and δ can be estimated by maximum likelihood. An alternative approach is to add the initial conditions as additional parameters for estimation, but this can lead to estimation problems. Forecasting simply continues the recursion with $\epsilon_t = 0$, keeping the parameters fixed.

AES includes the following forecasting methods:

[2]Makridakis, Spiliotis, and Assimakopoulos (2018) show the inferior forecasting performance of some machine learning and artificial intelligence methods on monthly M3 data.

SES	Simple exponential smoothing	$\delta = b_0 = 0$,
HES	Holt's exponential smoothing,	
Theta2	Theta(2) method	$\delta = 0, b_0 = \widehat{\tau}/2$ defined in (1).

The *SES* and *HES* models with infinite startup are ARIMA$(0,1,1)$ and ARIMA$(0,2,2)$ models respectively, see Hyndman, Koehler, Ord, and Snyder (2008, Ch.11). A dampened trend model adjusts the slope equation to $b_t = \phi b_{t-1} + \delta\epsilon_t$. Holt–Winter adds a seasonal equation to the system.

The Theta method of Assimakopoulos and Nikolopoulos (2003) first estimates a linear trend model

$$y_t = \mu + \tau(t-1) + u_t,\ t = 1, ..., T, \tag{1}$$

by OLS. The Theta forecasts are then the sum of the extrapolated trend and forecasts from the model for $y_t(\theta) = y_t - \widehat{\tau}(t-1)/\theta$, with weights $1/\theta$ and one respectively. The suggested model for $y_t(\theta)$ is SES, in which case this method can be implemented within the AES framework, as shown by Hyndman and Billah (2003). *Theta2*, i.e. using $\theta = 2$, had the best sMAPE (see §2.3 below) in the M3 competition, see Makridakis and Hibon (2000).

The AES estimates depend on several factors: initial conditions of the recursion, imposition of parameter constraints, and objective function. We have adopted different conventions for the initial conditions, so, in general, will get different results from Hyndman, O'Hara-Wild, Bergmeir, Razbash, and Wang (2017). The exception to this is *SES* with $0.001 \leq \alpha \leq 0.9999$ and l_0 as an estimated parameter. We also get almost identical results for *Theta2*, using $0.001 \leq \alpha \leq 0.9999$ and $l_0 = y_1 - b_0$, which conditions on the first observation to force $\epsilon_1 = 0$. For the yearly M3 data with $H = 6$ we obtain an sMAPE of 16.72, where Hyndman and Billah (2003, Table 1) report 16.62 (the submission to M3 has sMAPE 16.97).

2.2 Seasonality in the M4 benchmark methods

When a series has T observations with frequency $S > 1$, the seasonality decision in the M4 benchmarks is based on the Sth term in the ACF according to:

$$R(S) = T\frac{r_S^2}{1 + 2\sum_{j=1}^{S-1} r_i^2} \sim \chi^2(1). \tag{2}$$

If seasonality is detected with a p-value of 10% or less, it is estimated as the seasonal average of an $\text{MA}_{2\times S}$ filter (or just MA_S if S is odd). The benchmark method is then applied to the seasonally adjusted data.

The benchmark methods use multiplicative adjustment throughout. Assuming the frequency S is even, the seasonal component is the deviation from a smooth 'trend':

$$s_t = y_t/\text{MA}_{2\times S}(y_t), t = 1 + S/2, ..., T - S/2.$$

Now let $\bar{s}_{\cdot j}$ denote the average for each season from the $T - S$ observations s_t (so not all seasons need to have the same number of observations). These seasonal estimates are normalized to the frequency:

$$\widehat{s}_j = \bar{s}_{\cdot j} / \left[\frac{1}{S}\sum_{j=1}^{S} \bar{s}_{\cdot j}\right].$$

The seasonally adjusted series is

$$y_t^{\text{sa}} = y_{i,j}/\widehat{s}_j.$$

Finally, the forecasts from the seasonally adjusted series are multiplied by the appropriate seasonal factors.

2.3 M4 forecast evaluation

M4 uses two scoring measures, called MASE (Hyndman and Koehler, 2006) and sMAPE (Makridakis, 1993). For time-series $y_t, t = 1, ..., T+H$ with forecasts $\widehat{y}_t$ produced over $T+1, ..., T+H$ and seasonal frequency m:

$$\text{sMAPE} = \frac{100}{H} \sum_{t=T+1}^{T+H} \frac{|y_t - \widehat{y}_t|}{(|y_t| + |\widehat{y}_t|)/2}, \tag{3}$$

$$\text{MASE} = \frac{1}{H} \frac{\sum_{t=T+1}^{T+H} |y_t - \widehat{y}_t|}{\overline{|\Delta_m y|}}, \tag{4}$$

where the denominator of MASE is the average of the seasonal difference over the 'estimation period:'

$$\overline{|\Delta_m y|} = \frac{1}{T-m} \sum_{t=m+1}^{T} |y_t - y_{t-m}| .$$

MASE is infinite if the series is constant within each season: in that case we set it to zero.

The results below report the average MASE and sMAPE for each frequency. To facilitate comparison, these averages are often scaled by the average accuracy of the benchmark *Naive2* forecasts.

A $100(1-\alpha)\%$ forecast interval is expressed as $[\widehat{L}_t, \widehat{U}_t]$. Accuracy of each forecast interval for all series and horizons h is assessed on the basis of mean scaled interval score (MSIS):

$$\text{MSIS} = \frac{1}{\overline{\Delta_m y}} \frac{1}{H} \sum_{t=T+1}^{T+H} \left[\widehat{U}_t - \widehat{L}_t + \frac{2}{\alpha}(\widehat{L}_t - y_t) I(y_t < \widehat{L}_t) + \frac{2}{\alpha}(y_t - \widehat{U}_t) I(y_t > \widehat{U}_t) \right]. \tag{5}$$

M4 uses $\alpha = 0.05$, so that any amount outside the bands is penalized by forty times that amount. Gneiting and Raftery (2007, p.370) show that the interval score is 'proper,' meaning that it is optimized at the true quantiles.

We can also count the number of outcomes that are outside the given forecast interval. For a 95% pointwise interval we aim to be outside in about 5% of cases, corresponding to a coverage difference of close to zero.

3 Visualization of M4

Understanding the properties of such a large dataset is a challenge: it is too time consuming to look at each series separately. We found distribution plots as presented in Figure 1 useful. Each graph plots the frequency of outcomes along the vertical axis. To create Figure 1, we estimated $\widehat{\rho}$ by OLS from $\log y_t = \mu + \rho \log y_{t-1} + \epsilon_t$. Along the vertical axis in the first graph are the frequencies of $\widehat{\rho}$. The same data is shown in cumulative form along the horizontal axis. This dual perspective reveals the central tendency without hiding the tails. The estimates of ρ cluster near unity at all frequencies, corresponding to very high persistence in the data.

The top row of Figure 2 shows the distribution of the p-values of the test for a seasonal root (2), applied to the original y_t. This is the method used in the benchmarks, with the exception of daily data that we gave $S = 5$ to highlight the impact on inference. The null hypothesis assumes that there is no significant lower order serial correlation, which is mostly proven wrong by Figure 1. As a consequence, the incidence of seasonality is over-estimated, and a more accurate representation is to apply the test to $\Delta \log y_t$. This is shown in the middle row of graphs, now with a lower incidence of seasonality. This can be seen most strongly for daily data, which are tested with $S = 5$ but have no seasonality.

The bottom row of graphs shows the p-values of the ANOVA test for stable seasonality, applied to $\Delta \log y_t$ and as used in Doornik, Castle, and Hendry (2019).

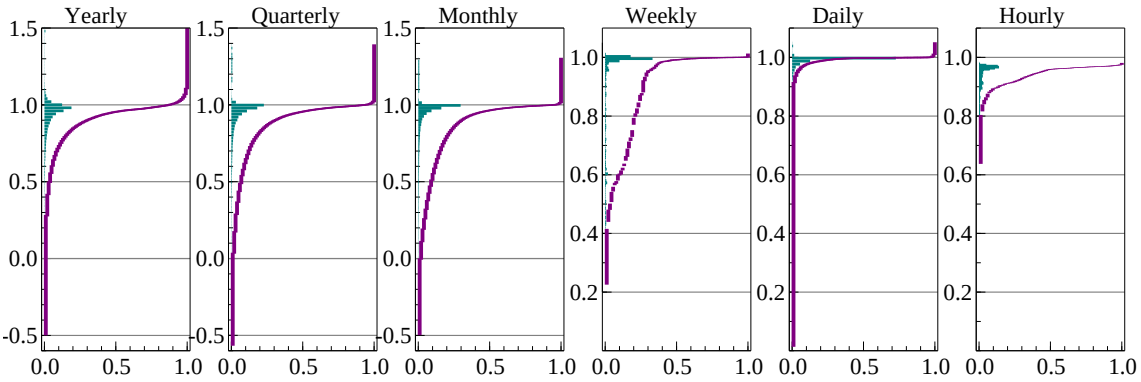

Figure 1: Distribution of autoregressive coefficient for each frequency of M4

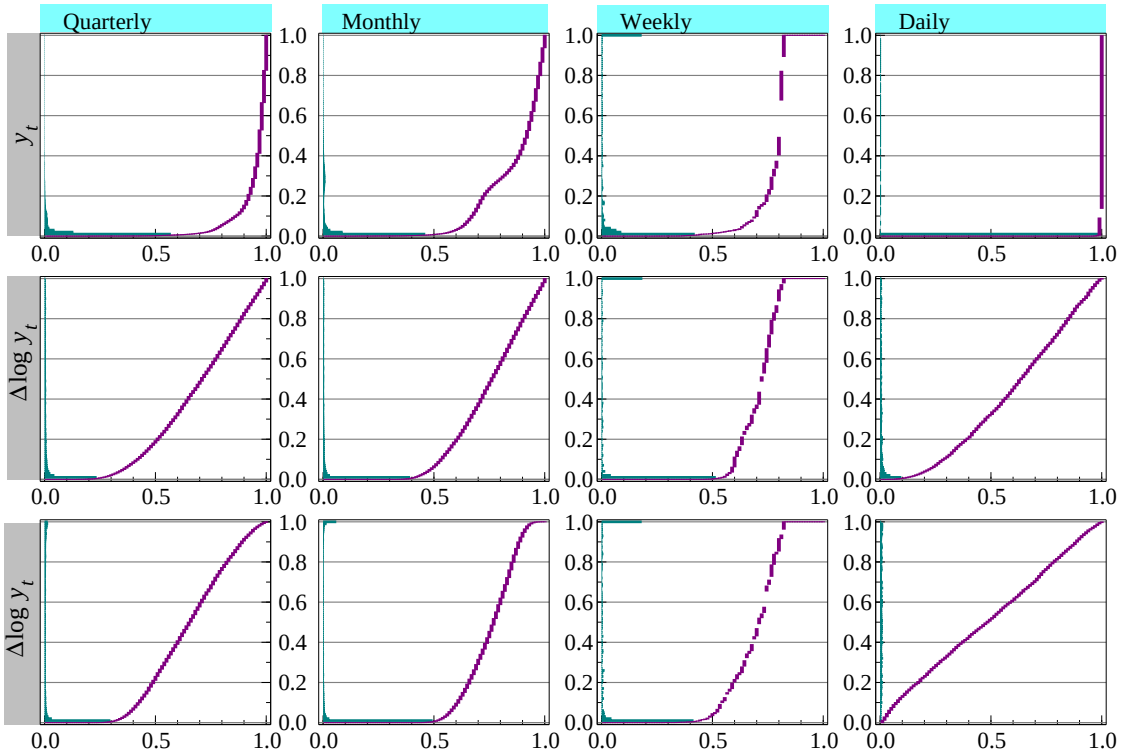

Figure 2: Tests of seasonality for quarterly, monthly, weekly and daily (with $S = 5$) M4. First row for y_t, second row for $\Delta \log y_t$, third row seasonal ANOVA test for $\Delta \log y_t$.

4 Evaluation of benchmark methods

4.1 An improved benchmark method: *Theta.log*

The benchmark methods do not take any transformations of the variables. In contrast, Bergmeir and Hyndman (2016) use a Box–Cox transformation in their bagging method. Legaki and Koutsouri (2018) improve on the Theta method in the M4 competition by using a Box–Cox transformation. In both cases the transformation is restricted to $\lambda \in [0, 1]$

$$y_t(\lambda) = \lambda^{-1} \left(y_t^{\lambda} - 1 \right),$$

where $\lambda = 0$ is the logarithmic transformation, and one is no transformation. The difference between the two approaches is that the former does the Box–Cox transformation before deseasonalization, and the latter afterwards. We found that this distinction matters little: the values of λ estimated before or

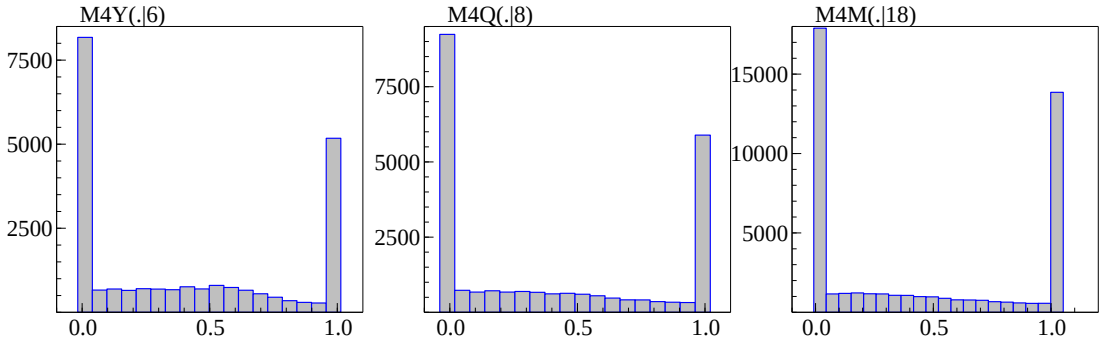

Figure 3: Estimated Box–Cox λ for yearly (left), quarterly (middle) and monthly M4 (right)

after multiplicative seasonal adjustments are similar: in quarterly M4 and monthly M3 the correlation between the estimates exceeds 0.9.

Figure 3 shows the histogram of λ estimated by maximum likelihood in a model on a constant and trend (and restricted to be between zero and one). This amounts to minimizing the adjusted variance as a function of λ. The U shapes in Figure 3 indicates that the choice is mainly between logarithms and levels. This suggests a simpler approach, such as comparing the variance when using levels to that using logs. Removing the trend by differencing leads to an approach as in Ermini and Hendry (2008), i.e. using logs when $\min y_t > 1$ in combination with:

$$\exp(2\overline{\log y})\text{var}[\Delta \log y_t] < c_l^2 \text{var}[\Delta y_t], \tag{6}$$

where $\text{var}[x_t]$ is the sample variance of $x_t, t = 1, ..., T$ and $\overline{\log y}$ is the sample mean of $\log y_t$. Using (6) means that the iterative estimation of λ can be avoided. All observations in M3 and M4 are positive, and experimentation suggests a benefit from preferring logs when $\widehat{\lambda} < 1$. The c_l^2 term is introduced to allow a bias towards using logarithms: based on forecast performance we adopt $c_l = 1.3$. As a consequence the proportion that is not logs in M4 is about 3% for annual data, 2% for quarterly and monthly, and less than 1% for the remainder.

4.2 A simplified benchmark method: *THIMA.log* and *THIMA*

To understand why Theta(2) is relatively successful, we introduce a simplified variant. Remember that SES is an ARIMA(0,1,1) model, and that the slope of the trend can also be estimated through the mean of the first differences. This leads us to suggest a trend-halved integrated moving average model (*THIMA*):

(1) Starting from $y_t,\ t = 1, ..., T$, the first differences $\Delta y_t,\ t = 2, ..., T$ have mean $\widetilde{\tau}$. Construct $x_t = \Delta y_t - \frac{1}{2}\widetilde{\tau}$.

(2) Estimate the following MA(1) model by nonlinear least squares (NLS) with $\widehat{\theta} \in [-0.95, 0.95]$:

$$x_t = \epsilon_t + \theta \epsilon_{t-1}$$

(3) The forecasts are:

$$\widehat{y}_{T+H} = y_T + \tfrac{1}{2}\widetilde{\tau} H + \widehat{\theta}\widehat{\epsilon}_T. \tag{7}$$

The forecasts from the MA(1) component are $\widehat{\theta}\widehat{\epsilon}_T$ for $H = 1$ and zero thereafter: their cumulation is constant (this is also the case for the SES forecasts). So *THIMA* forecasts consist of a dampened trend (arbitrarily halved), together with an intercept correction estimated by the moving average model. That estimation of one parameter does not have to be very precise, and the overall procedure is very fast. The *THIMA.log* version uses (6), again with $c_l = 1.3$, then is based on growth rates when the decision indicates the use of logs.

DGP	MASE	sMAPE
$y_t \sim \text{IN}[\mu, \sigma^2]$	1	$113\frac{\sigma}{\lvert\mu\rvert}$
$\Delta y_t \sim \text{IN}[\mu, \sigma^2]$	1	$\frac{200}{2T+1}\left[2\frac{\sigma}{\mu}\phi\left(\frac{-\mu}{\sigma}\right)+1-2\Phi\left(\frac{-\mu}{\sigma}\right)\right]$
$\Delta \log y_t \sim \text{IN}[\mu, \sigma^2]$	$\frac{m_3^T}{\frac{1}{T}\sum_{t=1}^{T} m_3^{t-1}}$	$200\frac{m_3-1}{m_3+1}$

Table 2: Approximate expectations of MASE and sMAPE under different data generation processes, $H = 1$. Φ is the standard normal cdf, ϕ the density, $m_3 = \exp(\mu + \sigma^2/2)$.

4.3 Evaluation

There is an extensive literature on forecast evaluation, and these measures will not give the same ranking of forecast performance. They are also sensitive to the adopted transformation of the variable, say differences versus levels (Clements and Hendry, 1993). In the context of M4, only MASE and sMAPE are relevant. The MASE is invariant to a change in location and scale of the target variable, while the sMAPE is invariant to rescaling y_t but not to a change in mean.

The sMAPE was introduced to address two issues with the MAPE (this is (3) but with just $|y_t|$ in the denominator): instability when outcomes close to zero are possible, as well as asymmetric response when exchanging outcomes and forecasts. However, it introduces two new problems. First, it favours overforecasting. As an illustration, take $y_t = \mu > 0, \mu > \delta > 0$ with forecast $\mu + \delta$ or $\mu - \delta$:

$$\text{sMAPE}(\mu + \delta) = \frac{2\delta}{2\mu + \delta} < \text{sMAPE}(\mu - \delta) = \frac{2\delta}{2\mu - \delta}.$$

This was already noted by Goodwin and Lawton (1999) and Koehler (2001). A second problem is that it introduces a bias that can be very large in some settings, illustrated in A.

For a better understanding, we derive approximate expectations of MASE and sMAPE when using random walk (naive) one-step forecasts. Three data generation processes (DGP) are considered: normally distributed in levels, stationary in differences and stationary growth rates. The results are derived in the Appendix, and summarized in Table 2. Interestingly, the approximate mean of MASE is unity unless the DGP is in logs. So this could be turned into another test of logs versus levels, complementing Spanos, Hendry, and Reade (2008). In the third case, the MASE is very roughly proportional to the sample size.

The sMAPE behaves very differently. In the white noise case it is a fixed multiple of the coefficient of variation (inverse signal-to-noise ratio). In the second case, difference stationary, the sMAPE is inversely proportional to the sample size. Finally, in the third case, it is largely independent of sample size again.

The top row of Figure 4 reports the average of the accuracy measures of the random walk forecasts for annual M3 and M4 in the line labelled *Naive2*. This is for $H = 1$ forecasts, so the first point of each line in each graph, when the horizontal axis is -12, twelve observations were withheld, eleven in the next, etc. The following table shows how the data are used:

	$1 \cdots\cdots T_i - H$	$T_i - H + 1 \cdots\cdots T_i$	$T_i + 1 \cdots\cdots T_i + H$
development	training	Test forecasts	unavailable
competition	competitor forecasts from this		M4 team tests
Fig. 4a, first	estimation	T	unused
Fig. 4a, second	estimation	T	unused
Fig. 4a, last	estimation		T

When developing our submission, we held back an additional H observations to test performance. In Figure 4 we use an expanding window, forecasting one observation ahead each time.

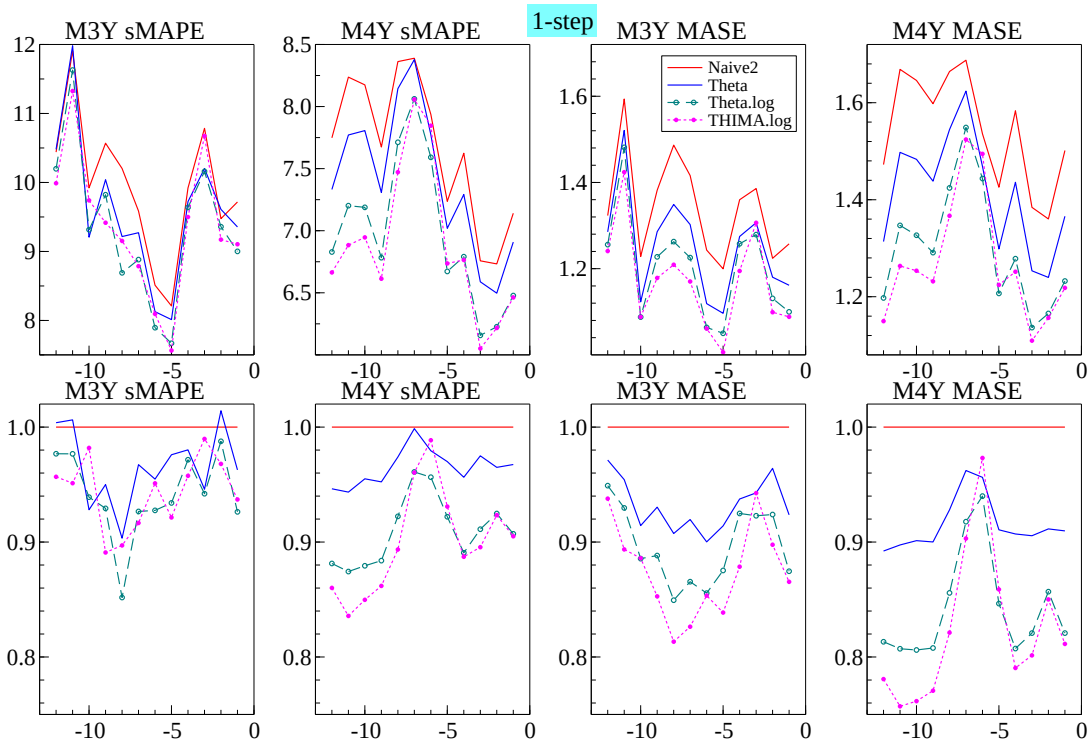

Figure 4: Average one step-ahead forecast accuracy for yearly M3 and M4, withholding from 12 to 1 observations at the end from the full dataset. The bottom row is normalized by the naive results.

First we note from the graphs that the variability is similar between M3 and M4, despite moving from 645 to 23 000 series for annual data. So uncertainty in rankings in M4 could be similar to that in M3. Next, accuracy rankings can switch for different subsamples, so it is not enough to claim success by looking at one particular sample. Finally, the M3 profile is somewhat U shaped, but M4 is more the opposite. This is relevant, because the competition version omits the last six observations, which corresponds to the middle of the graphs.

Comparing the approximations of 2 with the MASE in Figure 4, suggests that a representative DGP is that of $\Delta \log$. From simulation we find that $\mu = 0.025, \sigma = 0.1, T = 15$ gives sMAPE of 8.3 (this is a case where the approximation does not work) and MASE of 1.4. Standardizing by the naive results, as shown in the bottom row of Figure 4, reduces variability and makes the switch overs more visible.

Figure 4 also shows that all methods in the graph are, with a few exceptions, an improvement over the random walk forecast. As expected *THIMA.log* and *Theta.log* are similar. Table 3 reports the M3 performance of the improved and simplified benchmark methods, showing that both are improvements over the standard *Theta(2)* method. Their relative ranking is similarly unclear, because, dropping one observation at the end shows *THIMA.log* as the best performer. Seasonality is handled as in §2.2 in each case.

5 Adjustments to the *Card* method

Doornik, Castle, and Hendry (2019) provides a technical description of our *Card* method.

Experience has shown that a dampened trend is often useful for forecasting. Robustness is also helpful: the forecasts should not be thrown too much by a previous structural break. We introduce two forecasting methods that incorporate a dampened trend and robustness to structural breaks. The main focus is on autoregressive models, because these are easy to estimate and M4 provides no covariate

M3	Yearly($H=6$) sMAPE	MASE	Quarterly($H=8$) sMAPE	MASE	Monthly($H=18$) sMAPE	MASE
	Full sample, holdback H					
Theta.log	16.00	2.68	9.15	1.11	13.57	0.85
THIMA.log	16.10	2.68	9.19	1.11	13.75	0.86
Theta(2)	16.72	2.77	9.24	1.12	13.91	0.87
	With last observation removed, holdback H					
Theta.log	15.91	2.64	9.26	1.13	13.22	0.82
THIMA.log	15.61	2.57	9.07	1.10	13.22	0.82
Theta(2)	17.07	2.87	9.26	1.13	13.61	0.84

Table 3: M3 performance of MASE and sMAPE for *Theta(2)* and revised benchmark methods.

information. However, without precautions these can give quite wild forecasts in small samples.

Initial decisions are made about the use of logarithms, differences versus levels, and the presence of seasonality:

1. Let y_t denote the initial series. If $\min(y_1, ..., y_T) > 1$ take logs: $x_t = \log(y_t), t = 1, ..., T$, else $x_t = y_t$. This means that logs are always used in the M3 and M4 data.
   Forecasts for the logarithms are transformed back using $\widehat{y}_t = \exp(\widehat{x}_t)$ at the end (so not using a bias correction).
2. Compute sample variances of the differences and the levels. If $\text{var}[\Delta x_t] \leq 1.2\ \text{var}[x_t]$ then forecast from a dynamic model (if in differences, these must be cumulated to get level forecasts), else directly forecast the levels using a static model. The static model only occurs at a rate of 1.5% (yearly), 4% (quarterly), 6% (monthly), but almost never at the other frequencies.
3. The presence of seasonality is tested at 10% using the ANOVA test for stable seasonality, as used in Census X-11 seasonal adjustment (Ladiray and Quenneville, 2001). This is applied to Δx_t or x_t depending on the previous step. The bottom row of Figure 2, however, always used $\Delta \log y_t$, which is the most common case.

Our first forecast method is based on estimating the growth rates from first differences — hence labelled *Delta* method, but with removal of the largest values and additional dampening. This method uses means when levels forecasts are made.

The second method, called *Rho*, estimates a simple autoregressive model, possibly switching to a model in first differences with dampened mean.

Our final adjustment is to create a calibrated average of *Rho* and *Delta*, called *Card*. First, the forecasts of *Rho* and *Delta* are averaged with equal weights. Next is a calibration stage which treats the forecasts as if they were observed, and re-estimates a model that is a richer version of the first stage autoregressive model. The fitted values over the forecast period (now pseudo in sample) are the new forecasts. There is no issue with overfitting or explosive roots, because no further extrapolation is made.

Calibration makes little difference for annual or daily data, which have no seasonality. It does, however, provide almost uniform improvements in all other cases, in some cases substantially so. This experience is also reported for the X12-ARIMA procedure (Findley, Monsell, Bell, Otto, and Chen, 1998), although there the ARIMA model comes first, providing a forecast extension that is used in the X11 procedure. Our procedure could be a more flexible alternative.

Some further minor adjustments are made to allow for specific aspects at certain frequencies. For hourly data, the *Rho* and *Delta* are calibrated, then averaged, then calibrated again. Calibration is done with autoregressive lag six instead of one. For weekly data, *Rho* is applied to the four-weekly averages

(giving frequency 13), and calibrated before averaging with *Delta*.

In Doornik, Castle, and Hendry (2019) we specified the daily frequency as 5×12, but this was not beneficial, and we change it here to $S = S_2 = 1$. So no distinction is made anymore between yearly and daily data.

The Ox 7 (Doornik, 2013) code to replicate our Card submission was uploaded to Github shortly after the M4 competition deadline.

5.1 Robust forecasts

A correction can make the forecast more robust when there is an unmodelled break at the forecast origin. To illustrate, consider an autoregressive model of order one, AR(1):

$$y_t = \mu + \rho y_{t-1} + x_t'\beta + \epsilon_t, \quad t = 1, ..., T.$$

In the current setting, all components in x_t are deterministic, so known for the forecast period (see Castle, Doornik, and Hendry (2018) for an analysis where the future x_t's are not known). The one-step forecast is

$$\widehat{y}_{T+1} = \widehat{\mu} + \widehat{\rho} y_T + x_{T+1}'\widehat{\beta}.$$

The robust forecast is taken from the differenced model:

$$\widehat{y}^R_{T+1} = y_T + \widehat{\rho}\Delta y_T + \Delta x_{T+1}'\widehat{\beta} = \widehat{\mu} + \widehat{\rho} y_T + x_{T+1}'\widehat{\beta} + y_T - \widehat{\mu} - \widehat{\rho} y_{T-1} - x_T'\widehat{\beta} = \widehat{y}_{T+1} + \widehat{\epsilon}_T.$$

The robust forecast is an intercept correction based on the last residual. When nothing changes, $\mathsf{E}[\widehat{y}_{T+1}] = \mathsf{E}[\widehat{y}^R_{T+1}]$, but the variance is increased by $\widehat{\sigma}^2_\epsilon$. However, if there is a location shift in μ at T, this is captured in the residual, so there is a trade off between the increased variance and the reduced shift. A seasonal equivalent can be based on the seasonally differenced model:

$$\widehat{y}^{R(S)}_{T+1} = \widehat{y}_{T+1} + \widehat{\epsilon}_{T+1-S}.$$

Recursive application of robust forecasting leads to a rapid increase in the variance; for two steps ahead:

$$\widehat{y}^R_{T+2} = \widehat{y}^R_{T+1} + \widehat{\rho}\Delta\widehat{y}^R_{T+1} + \Delta x_{T+2}'\widehat{\beta} = (1+\widehat{\rho})\widehat{y}^R_{T+1} - \widehat{\rho} y_T + \Delta x_{T+2}'\widehat{\beta}.$$

In comparison the standard forecast is $\widehat{y}_{T+2} = \widehat{\mu} + \widehat{\rho}\widehat{y}_{T+1} + x_{T+2}'\widehat{\beta}$.

5.1.1 Robust adjustment for *Rho*

If the *Rho* model is not already estimated in differences (i.e. $I_\Delta = 0$), the following adjustment is made to $\widehat{x}_{T+1}$:

$$\begin{aligned} R &= \left[\tfrac{1}{2}(\widehat{\epsilon}_T + \widehat{\epsilon}_{T-S+1})\right]^{+2\widehat{\sigma}}_{-2\widehat{\sigma}}, \\ \widehat{x}^R_{T+1} &= \widehat{x}_{T+1} + \tfrac{1}{2}R. \end{aligned} \tag{8}$$

R is the averaged residual, which is limited to two residual standard errors, and half of that is added to the one-step forecast. So the robust forecast is the average of the original and the winsorized and shifted forecast. The notation $[x]^b_a$ indicates that x is bounded between a and b.

Figure 5 shows that the robust version of *Rho* using (8) leads to improvements at all frequencies, except for monthly and weekly data where the gains and losses are similar. The benefit is substantial for quarterly, daily and hourly data. It is small but consistent for annual data.

The results for M3 are similar: again there is no improvement for monthly data, although the annual improvement is more pronounced.

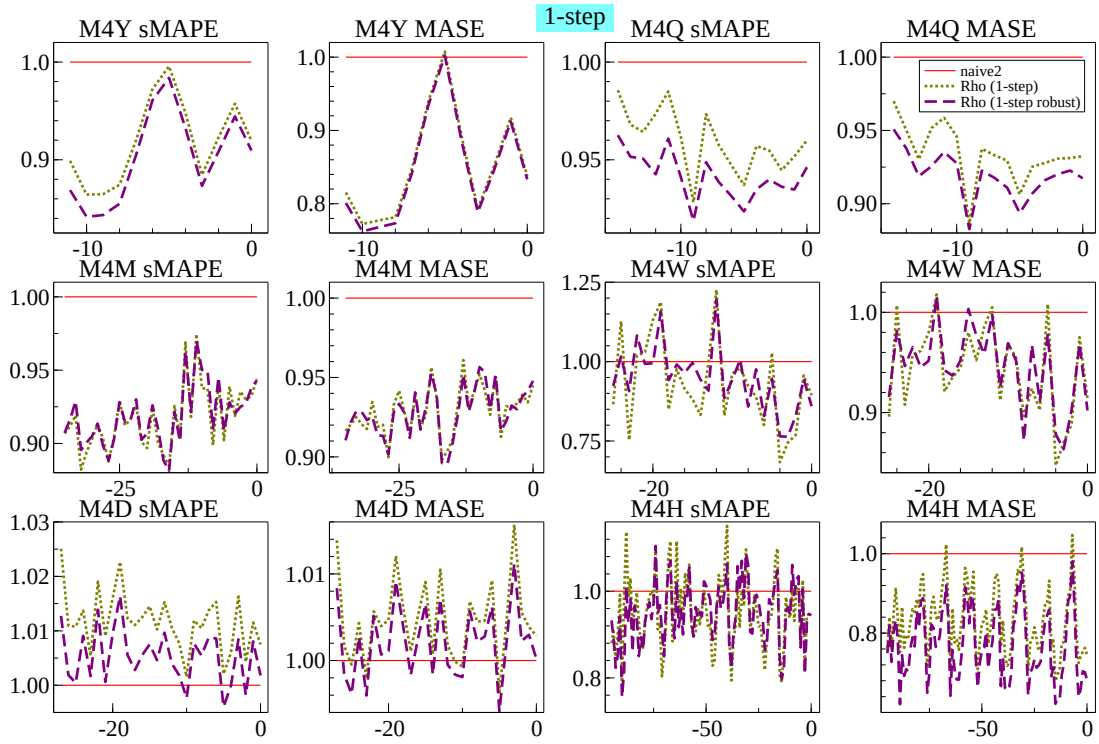

Figure 5: Accuracy of one-step forecasts from *Rho* and robustified *Rho* relative to *Naive2*. Recursive M4 data, all frequencies.

5.1.2 Robust adjustment for *Card*

At very short horizons, calibration performs worse than the inputs to calibration. We therefore made a small change for the first two forecasts, taking the average of the original and calibrated forecasts. Beyond $H = 3$, the forecasts are unchanged at the fitted values from calibration.

5.2 More averaging: *Cardt*

Our annual forecasts using *Card* did not do as well as expected. Part of the explanation is that holding back twelve observations from the full M4 data set is quite different from withholding six, as was illustrated in Figure 4. At the time we considered adding a Theta-like forecast to the average prior to calibration, but decided against this. That was a mistake, and we propose *Cardt* for frequencies up to twelve as the calibrated average of *Rho*, *Delta*, and *THIMA.log*.

5.3 Forecast intervals

The forecast intervals of an AR(1) model with drift, $y_t = \mu + \rho y_{t-1} + \epsilon_t$, grow with the horizon h when the errors are IID:

$$\text{SE} = \sigma \left(1 + \rho^2 + ... + \rho^{2(h-1)}\right)^{1/2}.$$

This is slower than the mean effect:

$$y_{T+h} = \mu \left(1 + \rho + ... + \rho^{h-1}\right) + \rho^h y_T.$$

Our submitted approach was based on $z_t = \log y_t$ with 90% forecast interval:

$$\exp\left[\widehat{z}_{T+h} \pm C_1 \widehat{\sigma} \left(1 + \widehat{\rho}_L + ... + \widehat{\rho}_L^{h-1}\right)\right],$$

where $\widehat{\rho}_L$ is an adjusted estimate of the autoregressive parameter and C_1 is determined by withholding data from the competition data set, aiming for a 90% interval. Even though this approach did well, it suffers from being asymptotically invalid, as well as being fixed at the 90% interval.

Our new approach makes a small-sample adjustment to the standard formula for the forecast interval. The basis for this is the calibration formula, which, however, is somewhat simplified and restrained. Because the sample size is small in some cases, we account for parameter uncertainty. Ideally, we get the correct point-wise coverage at each interval and in total, as well as a small MSIS.

The forecast bands have the following form:

$$\left[\widehat{L}_{T+h}, \widehat{U}_{T+h}\right] = \left[\exp\left(\widehat{z}_{T+h} \pm c_\alpha \left\{(\text{var}[\widehat{z}_{T+h}])^{1/2} + \frac{\pi_h(S)}{T}\right\}\right)\right]. \tag{9}$$

This is the standard forecast uncertainty for an autoregressive model with regressors, but here with an inflation factor π_h that depends on the frequency. The value of π_h is determined by withholding H observations from the training data, and finding a value that combined reasonably good coverage at all forecast horizons with a low value of MSIS. The form of π_h is given in B. The critical value c_α is from a student-t distribution. In addition, for $S = 4, 12, 52$, we average the forecasts standard errors from calibration in logs with those from calibration applied to the levels.

5.4 Evaluation

Figure 6 shows the performance of *Delta* and *Rho* at each fequency of M4. From $2H$ to H observations are withheld (see Table 1 for the value of H), so the last entry amounts to evaluating a submission to the competition. We see that *Delta* performs better than *Rho* for annual, weekly and daily data. Next, we consider *DelRho*, which is the simple average of the two. This always improves: it either matches the best of either (yearly, weekly, hourly) or improves considerably on both (quarterly, monthly). For daily data, it is very hard to beat the random walk, as expected from data that is largely financial. Finally, we consider the calibrated average. This improves a bit for quarterly monthly, but is a particularly effective way to handle the complex seasonal patterns of weekly and hourly data. Note that all plots in Figure 6 have the same scale, except for hourly data where the improvement over the naive forecast is so large.

Figure 7 looks at *THIMA.log*, *Card*, and *Cardt*. This shows that the addition of *THIMA.log* makes *Cardt* consistently outperform *Card*, except for daily (and weekly where it is not used). *THIMA.log* on its own is nowhere better, except for a short period with the yearly data.

Figure 8 shows the coverage of the forecast intervals, averaged up to H-steps ahead for $\alpha = 0.05$ and $\alpha = 0.1$. This shows that the forecast intervals are generally well behaved. The exceptions are that the 90% intervals for monthly data are a bit to wide, and hourly intervals a bit too narrow. The bands' effectiveness fluctuates with the subsample, perhaps more than expected.

These graphs average over all horizons, and are uninformative on any specific horizon. A good MSIS performance gives additional support for the adoped procedure. MSIS scores are reported in the bottom half of Table 4, which gives the summary statistics of our methods in the format that is used to determine M4 rankings. The new approach to computing forecast intervals is comparable in terms of coverage, but a considerable improvement as measured by MSIS. The new intervals have been derived after the competition finished — nonetheless, they are among the best methods in terms of MSIS.

Table 4 also gives the forecast performance in terms of MASE and sMAPE. As expected, the changes to *Card* have a negligable impact on its performance. The new *Cardt*, which adds *THIMA.log* to the combination, is mainly improved for annual data, and just a bit better for quarterly data.

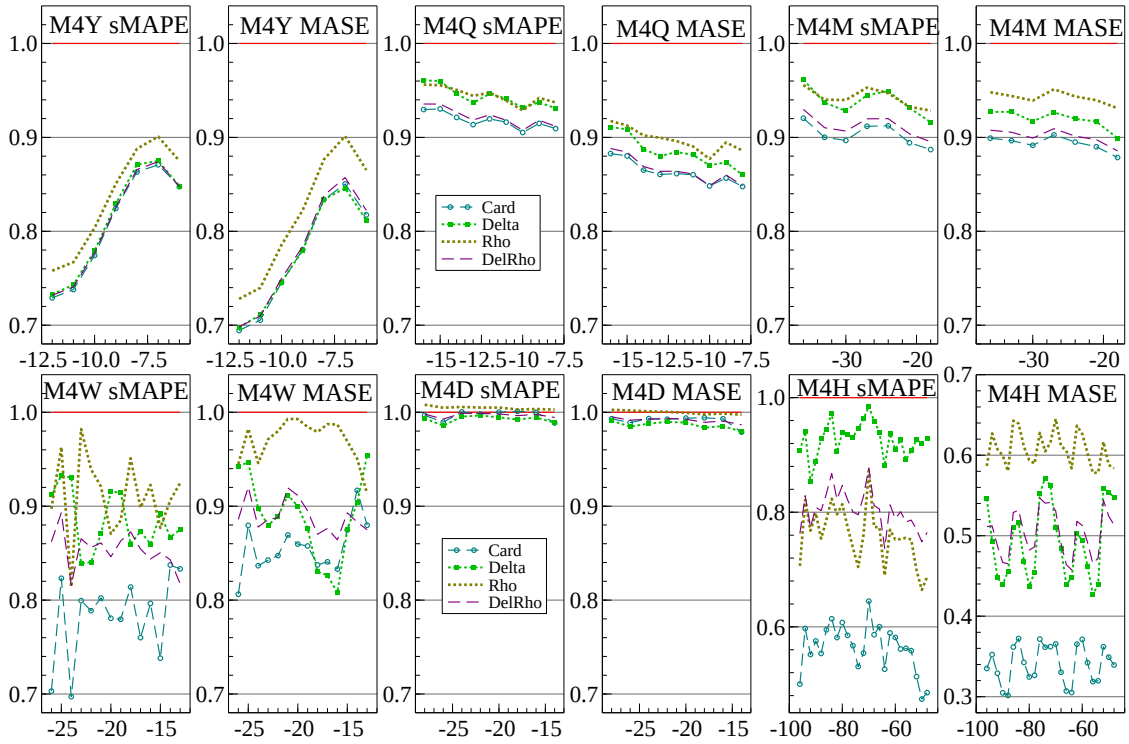

Figure 6: H-step forecast accuracy relative to that of *Naive2* for all frequencies of M4, retaining from $2H$ to H observations for evaluation. Forecast methods are *Delta*, *Rho*, $(Delta + Rho)/2$, *Card*.

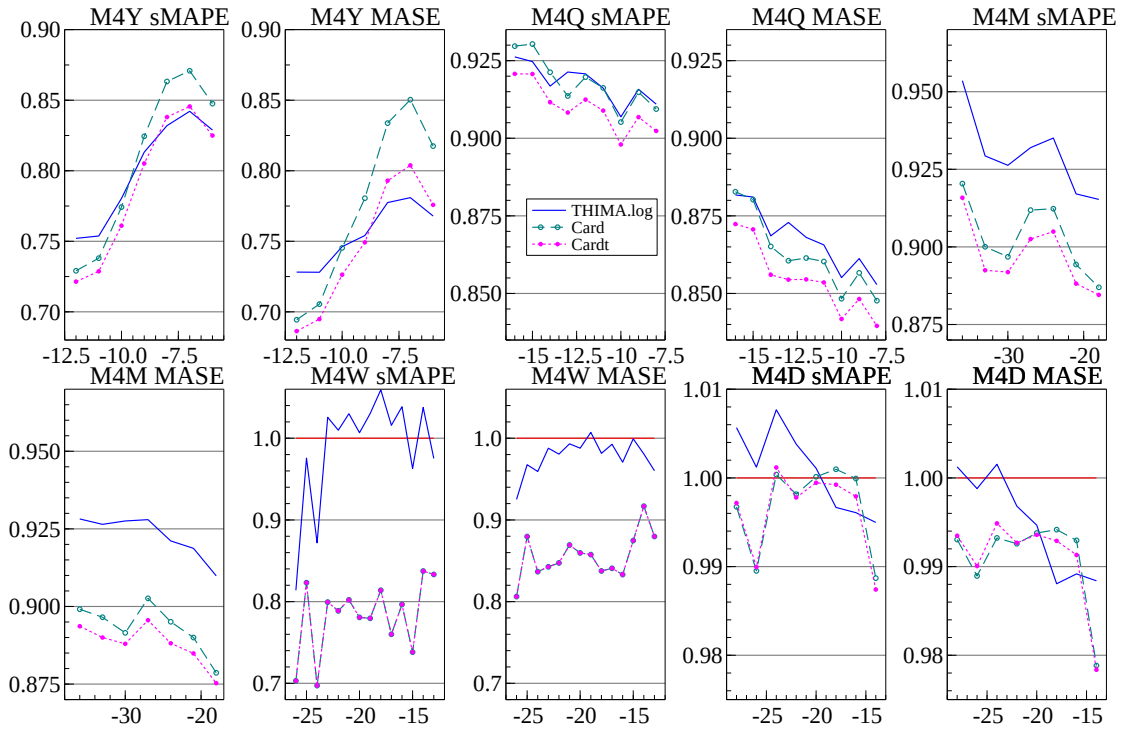

Figure 7: H-step forecast accuracy relative to that of *Naive2*. Forecast methods are *Card*, *Cardt*, *THIMA.log*.

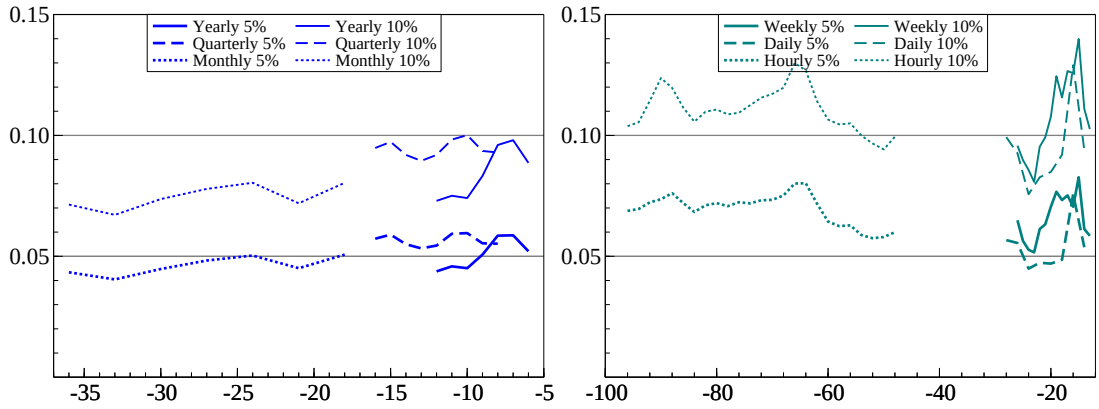

Figure 8: Average rejection of 95% and 90% H-step forecast intervals for all frequencies of M4, retaining from $2H$ to H observations for evaluation.

M4	Y	Q	M	W	D	H	Y	Q	M	W	D	H	All
	sMAPE						MASE						OWA
new *Cardt*	13.50	9.94	12.76	6.72	3.00	8.92	3.09	1.15	0.93	2.30	3.21	0.81	0.849
updated *Card*	13.87	10.02	12.80	6.72	3.01	8.92	3.26	1.16	0.93	2.30	3.21	0.81	0.864
submitted *Card*	13.91	10.00	12.78	6.73	3.05	8.91	3.26	1.16	0.93	2.30	3.28	0.80	0.865
new *THIMA.log*	13.55	10.03	13.21	7.91	3.02	18.41	3.05	1.17	0.97	2.55	3.23	2.50	0.865
	MSIS						ACD90%*/95%						MSIS	ACD
new *Cardt*	25.77	8.71	8.09	15.78	26.55	5.85	.002	.006	.005	.001	.006	.010	13.10	0.005
updated *Card*	26.50	8.82	8.12	15.78	26.60	5.85	.010	.009	.006	.001	.007	.010	13.31	0.008
submitted *Card**	30.20	9.85	9.49	16.47	29.13	6.14	.013	.021	.004	.003	.009	.048	15.18	0.007

Table 4: Summary performance in M4 competition. Absolute coverage difference is for a 95% forecast interval except for submitted *Card* which used 90%. OWA is the overall weighted average of sMAPE and MASE, with weights determined by the relative number of series for each frequency.

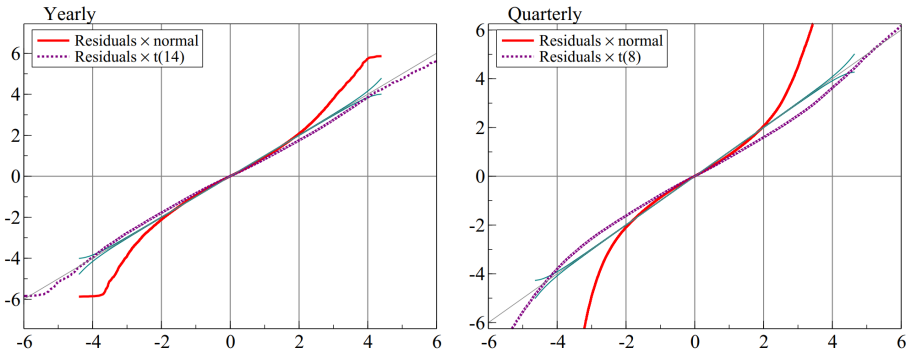

Figure 9: QQ plots of annual and quarterly residuals against Normal and closely matching Student-t distribution

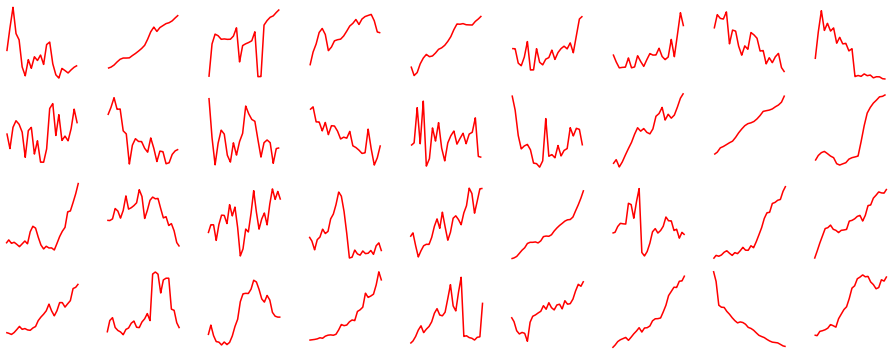

Figure 10: Eighteen actual yearly M4 series and eighteen simulated series

6 A simulation experiment

As a first step in designing an experiment that mimics some of the properties of M4, we address normality. The backbone of *Card* is the calibration method. We apply calibration to the yearly and quarterly series without forecasting, and collect the residuals, standardized by their estimated equation standard error. This gives 630 515 yearly and 2 010 696 quarterly residuals. Figure 9 shows that normality is strongly rejected: with so many residuals, the 95% error bands (see Engler and Nielsen, 2009) are very tight. Normality is matched well in the center between ± 2, but the residuals have fatter tails. This is not a surprise for economic data, where breaks happen intermittently. It also corresponds to the need to inflate the forecast intervals from calibration.

The following data generation process (DGP)

$$\begin{aligned} x_t &= \mu_0 U_1 + \rho x_{t-1} + \sigma\left[\delta_t + \epsilon_t + \theta\epsilon_{t-1}\right], \quad \epsilon_t \sim \text{IN}[0,1], \quad t = -99, ..., T \\ & U_1, U_2 \sim \text{IN}[1,1], \\ & \sigma = \sigma_0 + 0.02\left(U_2^2 - 1\right), \\ & \delta_t = 2u_t I(|u_t| > 2.58) \quad u_t \sim \text{IN}[0,1], t \geq 1(\delta_t = 0 \text{ for } t < 1), \\ y_t &= 100\exp(x_t - x_1), \end{aligned} \tag{10}$$

with parameters $\rho = 1, \theta = 0, \mu_0 = 0.03, \sigma_0 = 0.06$, closely matches the M4 yearly data in terms of the mean and standard deviation of growth rates, the distribution of the estimated autoregressive coefficient, as well as the shape of the QQ plot of the calibration residuals. The dominance of the unit root ($\rho = 1$) in logarithmic form was already established from Figure 1, the values of MASE and sMAPE, as well as the improved benchmark methods.

For a superficial comparison, we show in Figure 10 eighteen real M4 series, followed by the same number of simulated series. They look comparable, except that the third series would be unlikely to arise

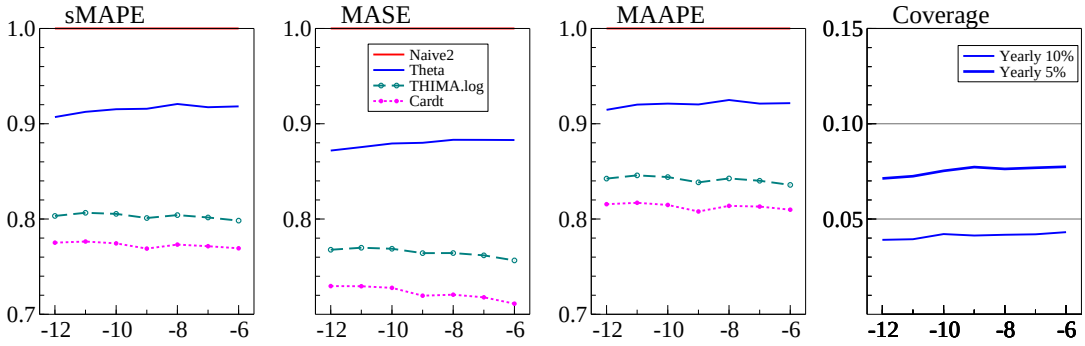

Figure 11: H-step forecast accuracy measured by average MASE, sMAPE, and MAAPE relative to that of *Naive2* for simulated data, retaining from $2H$ to H. 10 000 series.

from the DGP. The M4 data set has a few extremely large breaks that will not be replicated. Finally, the DGP could be lacking some heteroscedasticity.

However, the DGP (10) also has some advantages over M4. The first is that the series are independent (albeit highly correlated). This means that 'whole database' forecast methods do not inadvertently use future information, thus avoiding infeasible forecasts. Furthermore, generating data this way is much easier to implement, so can serve as an initial forecasting testbed. The DGP captures the properties that seem relevant for economic time series, which can help to improve machine learning methods in such a setting. Finally, there are some parameters that can be varied, in addition to the sample size.

Figure 11 evaluates forecasting six periods ahead, holding back from 12 to 6 observations from a sample of 32. The first two plots show the sMAPE and MAPE. In contrast to Figure 7, the results are much more stable over expanding windows. The third plot introduces the MAAPE, Kim and Kim (2016), which is defined as:

$$\text{MAAPE} = \frac{100}{H}\sum_{t=T+1}^{T+H} \text{atan2}\left(|y_t - \widehat{y}_t|, |y_t|\right).$$

The MAAPE avoids the problems of small values that MAPE and sMAPE have, and has a lower bias (see A). The final plot is the frequency of forecasts outside the 95% and 90% forecast intervals. The intervals are a bit too wide, showing that the inflation factor is overcorrecting.

Table 5 shows the forecast accuracy for data simulated from the DGP (10). The first three rows use default parameters ($\rho = 1, \theta = 0, \mu_0 = 0.03, \sigma_0 = 0.06$). The sample size $T = 28$ with six out-of-sample forecasts is representative for yearly M4. We also generate 250 observations (plus an extra six for evaluation). In the first case all 250 are used for forecasting, in the second case just the last 40, as is the default for *Cardt*. The presence of large shocks with probability of 1% means that there is no advantage here from using the whole data set. The bottom half of the table varies one parameter at a time. In all cases, *Cardt* outperforms *THIMA.log*, which in turn outperforms *Theta* (with one exception), although in some cases the difference is small. A larger μ_0 or smaller σ_0 increases the relative performance advantage of *Cardt*. Small changes in ρ have a large impact, but are not compatible with its observed distribution.

7 Conclusions

We established that the dominant features of M4 are mostly stationary growth rates that are subject to intermittent large shocks, combined with strong seasonality. This led us to propose a simple extension to the Theta method by adding a simple rule to take logarithms. We also introduced *THIMA.log* as a simple benchmark method that helps understanding the Theta method. Moreover, this improves on *Theta(2)* in both M3 and M4 forecasting at low frequencies. We added this to the forecast combination prior to

	T	MAAPE			sMAPE			MASE		
		Theta	*THL*	*Cardt*	*Theta*	*THL*	*Cardt*	*Theta*	*THL*	*Cardt*
		$\rho = 1, \theta = 0, \mu_0 = 0.03, \sigma_0 = 0.06$								
Default DGP	28	14.73	13.36	12.94	16.56	14.40	13.87	4.00	3.42	3.22
Default DGP	250/250	22.34	20.02	18.49	28.20	24.64	22.51	25.68	18.40	15.90
Default DGP	250/40	15.12	13.60	13.03	17.12	14.81	14.00	23.26	18.42	16.62
Default DGP	28	14.73	13.36	12.94	16.56	14.40	13.87	4.00	3.42	3.22
$\sigma_0 = 0.03$	28	11.84	9.92	8.58	13.22	10.64	9.15	4.09	3.21	2.58
$\sigma_0 = 0.09$	28	18.20	17.23	17.12	20.69	18.66	18.51	3.93	3.57	3.50
$\mu_0 = 0.01$	28	12.63	12.46	12.39	13.55	13.01	12.96	2.46	2.44	2.43
$\mu_0 = 0.05$	28	18.54	14.94	13.49	22.37	16.60	14.69	6.44	4.82	4.05
$\theta = 0.2$	28	15.79	14.45	14.13	17.90	15.61	15.18	4.25	3.69	3.52
$\theta = -0.2$	28	13.86	12.41	11.94	15.48	13.37	12.79	3.72	3.14	2.91
$\rho = 1.01$	28	30.82	22.58	17.85	44.31	28.01	20.78	19.17	13.86	10.20
$\rho = 0.98$	28	12.58	12.51	12.30	13.20	12.86	12.66	2.11	2.13	2.09
$\rho = 0.90$	28	12.27	12.29	12.04	12.64	12.59	12.32	1.89	1.90	1.86

Table 5: Summary performance in DGP. *THL* is short for *THIMA.log*. T is the sample size, with $H = 6$ out-of-sample observations for evaluation. $T = 250/40$ means that 250 observations are available, but only the last 40 used for forecasting. 10 000 series.

calibration, which mainly improved performance at the yearly frequency.

Our experience with M4 supports most of the principles that were introduced in the introduction:

A] dampen trends/growth rates;
This certainly holds for our methods and Theta-like methods. Both *Delta* and *Rho* explicitly squash the growth rates. *Theta(2)* halves the trend. The *THIMA* method that we introduced halves the mean of the differences, which has the same effect.

B] average across forecasts from 'non-poisonous' methods;
This principle, which goes back to Bates and Granger (1969), is strongly supported by our results, as well as the successful methods in M4. There may be some scope for clever weighting schemes for the combination, as used in some M4 submissions that did well. It may be that a judicious few would be better than using very many.
A small amount of averaging also helped with forecast intervals, although the intervals from annual data in levels turned out to be 'poisonous.'

C] include forecasts from robust devices in that average;
We showed that short-horizon forecasts of *Rho* could be improved by overdifferencing when using levels. The differenced method already has some robustness, because it reintegrates from the last observation. This, in turn, could be an adjustment that is somewhat too large. The IMA model of the *THIMA* method effectively estimates an intercept correction, so has this robustness property (as does *Theta(2)*, which estimates it by exponential smoothing).

D] select variables in forecasting models at a loose significance;
Some experimentation showed that the seasonality decisions work best at 10%, in line with this principle. Subsequent pruning of seasonal dummies in the calibration model does not seem to do much, probably because we already conditioned on the presence of seasonality. However, for forecast uncertainty, a stricter selection helps to avoid underestimating the residual variance. Castle, Doornik, and Hendry (2018) find support for this in a theoretical analysis.

E] update estimates as data arrive, especially after forecast failure.
This aspect was not covered here.

We derived a DGP that generates data similar to annual M4. This could be a useful complement to the actual data. It also confirmed the good performance of our *Cardt* method. Extending this to include the properties of the seasonal time series is left to a later date. Another possible refinement is to consider whether the data categories (macro, micro, etc.) have different time-series properties.

A Accuracy measures under naive forecasts

The random walk (or naive) forecast of annual data is simply $\widehat{y}_{T+1} = y_T$, so for one step ahead MASE, from (4):

$$\text{MASE(N1)} = \frac{|\Delta y_{T+1}|}{\frac{1}{T}\sum_{t=1}^{T}|\Delta y_t|}. \tag{11}$$

Similarly for MAPE and sMAPE:

$$\text{MAPE(N1)} = 100\frac{|\Delta y_{T+1}|}{|y_{T+1}|}, \tag{12}$$

$$\text{sMAPE(N1)} = 200\frac{|\Delta y_{T+1}|}{|y_{T+1}| + |y_T|}. \tag{13}$$

In this setting we can derive approximate expected values of these measures.

A.1 Stationary case, I

First assume that the data generation process is given by:

$$y_t = \mu + \epsilon_t, \quad \epsilon_t \sim \text{N}[0, \sigma^2].$$

Then $\Delta y_t \sim \text{N}[0, 2\sigma^2]$, therefore $|\Delta y_t|$ has a half normal distribution, and

$$\text{E}[|\Delta y_t|] = \sigma\left(\frac{4}{\pi}\right)^{1/2} = 2\sigma\phi(0) \equiv m_1.$$

Using the first term of the Taylor expansion around the expectation amounts to approximating the expectation of the ratio by the ratio of the expectations:

$$\text{E[MASE(N1)]} \approx \frac{\text{E}\left[|\Delta y_{T+1}|\right]}{\frac{1}{T}\sum_{t=1}^{T}\text{E}[|\Delta y_t|]} \approx 1.$$

The numerator and denominator of the MASE(N1) in (11) are asymptotically uncorrelated because the distributions of $|\Delta y_t|$ and $|\Delta y_s|$ are independent for $s < t - 1$.

For the denominator of sMAPE, note that $|y_t|$ has a folded normal distribution:

$$\text{E}[|y_t|] = 2\sigma\phi\left(\frac{-\mu}{\sigma}\right) + \mu\left[1 - 2\Phi\left(\frac{-\mu}{\sigma}\right)\right] \equiv m_2.$$

So

$$\text{E[sMAPE(N1)]} \approx 100\frac{m_1}{m_2}.$$

When μ/σ is large enough:

$$\text{E[sMAPE(N1)]} \approx 100\frac{m_1}{\mu} = 113\frac{\sigma}{|\mu|}.$$

This does not hold when $\mu = 0$: in that case the expectation is approximately $100\sqrt{2} = 141$. The numerator and denominator of the sMAPE (13) are uncorrelated provided y_{T+1} and y_T have the same sign (an alternative version with $|y_{T+1} + y_T|$ in the denominator would always be uncorrelated).

	$\mu = 0.5$	$\mu = 1$	$\mu = 2$	$\mu = 5$	$\mu = 10$
			Simulated		
E[MASE]	1.136	1.136	1.136	1.136	1.136
E[sMAPE]	133.9	109.1	63.2	23.0	11.3
Bias[MASE]	−0.006	−0.006	−0.006	−0.006	−0.006
Bias[sMAPE]	1.100	0.923	0.573	0.198	0.093
Bias[MAPE]	−0.279	−0.932	−1.640	−0.222	−0.108
Bias[MAAPE]	0.064	0.038	−0.109	−0.159	−0.100
			Approximated		
E[MASE]	1	1	1	1	1
E[sMAPE]	225.7	112.8	56.4	22.6	11.3
Bias[MASE]	0	0	0	0	0
Bias[sMAPE]	2	1	0.5	0.2	0.1

Table 6: Simulated and approximated mean and bias of MASE and sMAPE for one-step ahead naive forecasts. DGP N[μ, 1], $T = 15$, $M = 100\,000$ replications.

The approximations can provide some inside in the amount of bias introduced when minimizing the error measures: what is the optimal amount β to add to the naive forecast in the current DGP. In the MASE, only the nominator is affected, turning it into a folded normal distribution. Then minimizing the approximation amounts to minimizing $f(\beta) = 2\phi(-\beta) + \beta[1 - 2\Phi(-\beta)]$. Because $\partial f(\beta)/\partial\beta = 1 - 2\Phi(-\beta)$, this is zero at $\beta = 0$: the bias comes from the higher order terms that were ignored in the approximation.

In case of the sMAPE the bias function is more difficult because β also enters the denominator. For $\mu > 0$, sMAPE is roughly minimized for $\beta = \sigma^2/\mu$. Table 6 compares the approximations to simulations for a range of means when the variance equals one, confirming the increasing accuracy of the approximate expectations as μ increases.

It also shows that while the MASE is essentially unbiased, the bias from minimizing sMAPE and MAPE is large in some cases. The MAAPE was recently introduced by Kim and Kim (2016), and is defined as:

$$\text{MAAPE} = \frac{100}{H}\sum_{t=T+1}^{T+H} \text{atan2}\left(|y_t - \widehat{y}_t|, |y_t|\right).$$

A.2 Nonstationary case, II

Keeping T fixed:

$$\Delta y_t = \mu + \epsilon_t, \quad \epsilon_t \sim \text{N}[0, \sigma^2], t = 1, ..., T + 1.$$

Then $|\Delta y_t|$ has a folded normal distribution with mean m_2 and E[MASE] ≈ 1.

Setting $y_0 = 0$ we have that $y_t = \sum_{s=1}^{t} \Delta y_s \sim \text{N}[t\mu, t\sigma^2]$.

$$\text{E[sMAPE(N1)]} \approx 200\frac{m_2}{(2T+1)\mu}.$$

This is roughly $100/T$ when μ/σ is large, but more like $(100/T)\sigma/\mu$ for small μ.

A.3 Nonstationary case, III

Let

$$\begin{aligned} y_t &= \exp(x_t), \\ \Delta x_t &= \mu + \epsilon_t, \quad \epsilon_t \sim \mathrm{N}[0, \sigma^2], \\ x_t &= \sum_{s=1}^{t} \Delta x_t, \quad x_0 = 0. \end{aligned}$$

For fixed T, $\exp(\Delta x_t)$ has a lognormal distribution with mean $\exp(\mu + \sigma^2/2) \equiv m_3$, and $y_t = \exp(x_t)$ has a lognormal distribution with mean $\exp(t\mu + t\sigma^2/2) = m_3^t$, so

$$\mathrm{E}[\Delta y_t] = \mathrm{E}[[\exp(x_t)] - \mathrm{E}[[\exp(x_{t-1})] = m_3^{t-1}(m_3 - 1).$$

Ignoring the absolute values:

$$\mathrm{E}[\mathrm{MASE}] \approx \frac{m_3^T}{\frac{1}{T}\sum_{t=1}^{T} m_3^{t-1}}.$$

This is always positive, because $m_3 - 1$ cancelled out in this approximation. As a consequence, the approximation remains somewhat effective even for negative μ. Note that the MASE tends to zero as μ gets more negative. For larger μ the MASE behaves as Tm_3.

Finally, for the sMAPE when μ is large:

$$\mathrm{E}[\mathrm{sMAPE(N1)}] \approx 200\frac{m_3^{T+1} - m_3^T}{m_3^{T+1} + m_3^T} = 200\frac{m_3 - 1}{m_3 + 1}.$$

B Forecast intervals

Forecast intervals are obtained from the calibration model that is used to create the final forecasts. The calibration model is:

$$\begin{aligned} z_t = \mu &+ (\rho z_{t-1} + \rho_R z_{t-R} I_R I_4 I_S + \rho_{R+1} z_{t-R-1} I_R I_4 I_S) I_\rho + \{\delta_j q_{j,t}\} I_A I_S \\ &+ (\gamma_1 S_t + \gamma_1^* C_t)(1 - I_A) I_S + (\gamma_2 S_{2,t} + \gamma_2^* C_{2,t})(1 - I_3) I_{S2} \\ &+ \rho_{SS_2} z_{t-SS_2} I_3 I_{S2} + (\tau_1 d_t + \tau_2 t d_t I_5 I_\rho) I_6 + u_t, \quad t = T_0, ..., T + H \end{aligned} \tag{14}$$

where $I_\rho = 0$ for a static model, $I_S = S > 1$, $I_R = 1$ when $R > 1$, $I_4 = T > 4S$, $I_3 = T + H > 4SS_2$, $I_{S2} = S_2 > 1$, $I_6 = S{\neq}24$ *and* $T > 3S$ *and* $T + H - k > 10$, $I_5 = S{=}4, 12, 13$, $S_t = \sin[2\pi t/S], C_t = \cos[2\pi t/S]$, $S_{2,t} = \sin[2\pi t/(SS_2)], C_{2,t} = \cos[2\pi t/(SS_2)]$, $d_t = I(t < T - \min[2S, (T + H)/2])$. Note that no observations are lost when lag SS_2 is used, because the first SS_2 observations are duplicated at the start.

The preliminary forecasts are treated as if they are insample observations, and then replace by fitted values from calibration. However, the forecast error variance can only be estimated from out-of-sample extrapolation. This makes it essential to avoid explosive behaviour; we also wish to avoid underestimating the residual variance, so (14) is adjusted as follows:

1. remove the broken intercept and trend (if present, so setting $I_6 = 0$);
2. remove deterministic variables that are insignificant at 2%; the intercept and autoregressive part are not changed;
3. add the absolute residuals from (14) as a regressor;
4. estimate the reformulated calibration model;

	$y_t \sim \mathrm{N}[\mu, \sigma^2]$		$\Delta y_t \sim \mathrm{N}[\mu, \sigma^2]$		$\Delta \log y_t \sim \mathrm{N}[\mu, \sigma^2]$	
	mean	sdev	mean	sdev	mean	sdev
			$\mu = 0, \sigma = 1, T = 15$			
MASE	1.14	0.92	1.04	0.84	2.9	9.8
sMAPE	144.1	68.5	51.0	58.9	69.9	45.4
MAPE	6629.4	$18{\times}10^5$	809.6	$1.6{\times}10^5$	112.4	195.1
MAAPE	90.8	40.1	40.3	38.8	58.7	37.9
			$\mu = 0.025, \sigma = 0.1, T = 15$			
MASE	1.14	0.92	1.05	0.85	1.4	1.2
sMAPE	141.3	69.3	37.7	50.0	8.3	6.3
MAPE	654.0	10871	365.7	15478	8.5	6.8
MAAPE	89.5	40.6	31.7	34.8	8.4	6.7
			$\mu = 0.1, \sigma = 1, T = 15$			
MASE	1.14	0.92	1.04	0.84	4.3	13.2
sMAPE	143.8	68.6	48.5	57.6	70.1	45.5
MAPE	797.4	23462	188.7	6304.6	124.6	219.1
MAAPE	90.6	40.2	38.6	38.2	60.8	39.6
			$\mu = 1, \sigma = 1, T = 15$			
MASE	1.14	0.92	1.04	0.75	41.8	71.6
sMAPE	109.1	71.0	8.60	7.02	94.0	51.6
MAPE	648.1	27743	9.31	21.5	358.8	579.4
MAAPE	75.3	43.5	9.07	7.52	91.9	46.8
			$\mu = 10, \sigma = 1, T = 15$			
MASE	1.14	0.92	1.00	0.104	$5.1{\times}10^5$	$6.6{\times}10^5$
sMAPE	11.3	8.63	6.90	0.69	200.0	0.039
MAPE	11.5	9.05	7.15	0.74	$36{\times}10^5$	$47{\times}10^5$
MAAPE	11.3	8.71	7.14	0.74	157.1	0.010

Table 7: Simulated means and standard deviations of MASE, sMAPE, MAPE, and MAAPE for one-step ahead naive forecasts. $T = 15$, $M = 100\,000$ replications.

5. if $\widehat{\rho} > 0.999$ then impose the unit root, and re-estimate;
6. if $\widehat{\rho} < 0$ then set $\rho = 0$, and re-estimate.

Let $\widehat{u}_t$ denote the reformulated calibration residuals, then the equation standard error is estimated from 'recent' residuals:

$$\widetilde{\sigma}_u^2 = \sum_{\max(T-T^*+1,T_0)}^{T} \frac{\widehat{u}_t^2}{\max[\min(T^*, T - T_0 + 1 - k^*), 2]},\ T^* = \max(SS_2, 80), \tag{15}$$

where k^* is the number of regressors in the reformulated calibration model. The variance (15) is computed from 'recent' residuals to reflect neglected (conditional) heteroscedasticity. T is the original sample size excluding the forecast period, so the residuals from the forecast period are excluded; T_0 equals 1 for a static model, 2 for a model with one lag, and $R + 2$ if the seasonal lag is included.

The parameters of the adjusted model are estimated using all observations, but the forecast variance is extrapolated using the standard autoregressive forecast formulae from $T + 1$ onwards. This can be represented as

$$\widehat{\sigma}_u^2 \left[\widehat{f}_{T+h}^u + \widehat{f}_{T+h}^x \right],$$

where f^u is the contribution from the error term, and f^x the contribution of parameter estimation, with the former dominating asymptotically.

Two small adjustments are made: we use the recent residual variance (15) and limit the contribution from parameter uncertainty:

$$\text{var}[\widehat{z}_{T+h}] = \widetilde{\sigma}_u^2 \left[\widehat{f}^u_{T+h} + \min(\widehat{f}^x_{T+h}, 4\widehat{f}^u_{T+h}) \right].$$

By default, the modelling is in logs, so that the $100(1-\alpha)\%$ interval is given by:

$$\left[\widehat{L}_{T+h}, \widehat{U}_{T+h}\right] = \left[\exp\left(\widehat{z}_{T+h} \pm c_\alpha \left\{ (\text{var}[\widehat{z}_{T+h}])^{1/2} + \frac{\pi_h(S)}{T} \right\}\right)\right]. \tag{16}$$

The critical value c_α is from a student-t distribution with $T - T_0 - k^*$ degrees of freedom.

The π_h term is an inflation factor for $S = 4, 12, 24$ that is added when using logarithms because otherwise the forecast intervals are too small, particularly at longer horizons:

$$\pi_h(S) \begin{cases} 0.25h & S = 1, \\ 0.1(h-1) & S = 4, \\ 0.4h & S = 12 \\ 0.4\lfloor h/6 \rfloor & S = 24 \\ 0.0 & S = 52 \end{cases}$$

One further step for forecast intervals from calibation in logarithms is to also calibrate the levels, and then average the two. In that case the forecast standard errors are multiplied by $1 + 4h/T$. This is used for $S = 4, 12, 52$. For yearly data the levels forecast intervals are too far out to be useful in a combination.

C Some more results

Table 8 presents some additional performance comparisons of different methods. The top half of the table looks at different forecast horizons for the annual data using the competition data set (so excluding the evaluation data): in the first row, one observation is withheld for forecasting, in the second row, two, etc. In this example, the relative forecast performance of all methods gets increasingly better than the random walk forecast as the horizon grows.

The next block of Table 8 consideres several transformations. Because the implicit null hypothesis is that the growth rates are approximately normal, we may find preformance for transformations quite different. Note that in this case the MASE and sMAPE are expressed in terms of the transformed variables. Random1 draws from a normal distribution with the same mean and variance as the growth rates of the original series, then reintegrates:

$$\exp\left\{1 + \sum_{s=1}^{t} u_s\right\}, \quad u_t \sim \text{N}\left[\mu = \overline{\Delta \log y_t}, \sigma^2 = \text{var}(\Delta \log y_t)\right].$$

Random2 is similar to a wild bootstrap:

$$y_1 \exp\left\{\sum_{s=2}^{t} \mu + (\Delta \log y_s - \mu)\epsilon_s\right\}, \quad \epsilon_t \sim \text{N}\left[0, 1\right].$$

The results from the range of transformations shows that, in terms of sMAPE, *THIMA.log* always improves over *Theta2*, and *Cardt* is better again, occasionally by a large amount. The same broadly holds for MASE as well. For y_t^{-1} all methods are worse than *Naive2* on sMAPE, but it is likely that it is ill-defined in that case, with observations close to zero.

Cardt usually improves on the best of *THIMA.log*, *Rho* and *Delta*.

		sMAPE					MASE				
	H	*Theta*	*THL*	*Delta*	*Rho*	*Cardt*	*Theta*	*THL*	*Delta*	*Rho*	*Cardt*
		Yearly M4 relative to *Naive2*									
last observation	1	1.00	0.96	0.96	0.96	0.94	0.96	0.90	0.92	0.93	0.90
last two	2	0.96	0.89	0.91	0.91	0.89	0.92	0.83	0.86	0.86	0.83
last three	3	0.93	0.85	0.84	0.85	0.83	0.89	0.78	0.77	0.79	0.75
last four	4	0.91	0.80	0.78	0.80	0.77	0.89	0.77	0.75	0.77	0.74
last six	6	0.88	0.75	0.73	0.76	0.72	0.86	0.73	0.70	0.73	0.69
		Transformed yearly M4 data relative to Naive2									
y_t	6	0.88	0.75	0.73	0.76	0.72	0.86	0.73	0.70	0.73	0.69
$\Delta\Delta y_t$	6	1.10	1.04	1.13	1.02	1.05	0.83	0.87	0.85	0.88	0.86
Δy_t	6	0.97	0.97	0.99	0.97	0.97	0.92	0.95	0.95	0.94	0.94
$\Delta \log$	6	0.96	0.96	0.97	0.99	0.97	0.91	0.93	0.92	0.93	0.91
$\log y_t$	6	0.75	0.73	0.75	0.79	0.72	0.74	0.71	0.71	0.76	0.70
$\sum_{s=1}^{t} y_s$	6	0.66	0.53	0.37	0.69	0.23	0.72	0.58	0.38	0.85	0.25
$y_t^{1/2}$	6	0.82	0.75	0.74	0.76	0.72	0.79	0.71	0.68	0.71	0.67
y_t^2	6	0.96	0.75	0.73	0.76	0.72	1.00	1.00	1.00	1.00	1.00
y_t^{-1}	6	1.69	1.76	1.11	1.89	1.46	0.94	0.94	0.85	1.04	0.89
Random1	6	0.92	0.81	0.72	0.72	0.73	0.86	0.71	0.61	0.63	0.63
Random2	6	0.92	0.76	0.64	0.68	0.68	0.87	0.73	0.60	0.64	0.64
y_{T-t+1}	6	1.19	0.81	0.70	0.75	0.73	0.82	0.81	0.76	0.81	0.77

Table 8: Forecast comparison for yearly M4 training data relative to *Naive2* for *Theta2*, *THIMA.log*, *Delta*, *Rho*, and *Cardt*.

D Comparison with R

The table compares *Theta2* forecasts using the R code supplied with the M4 competition to our Ox implementation (using all data). The forecast summary statistics are almost identical except for weekly data, where we get a different result. The Ox version is about 130 times faster. Of that advantage, a factor of three is obtained from the parallel implementation.

Theta2	Ox implementation		R implementation		
	Time (sec)	sMAPE	Time (seconds)		sMAPE
	Total		Data	Forecast	
Yearly	3.67	0.876	4.61	375	0.880
Quarterly	4.83	0.949	4.31	627	0.950
Monthly	10.49	1.017	4.22	1499	1.016
Weekly	0.45	0.838	3.95	33	0.886
Daily	6.96	1.007	4.05	1019	1.008
Hourly	0.27	0.991	3.89	28	0.991

E Comparison with 118

Method 118 by Smyl (2018) had the best performance for yearly, quarterly, and monthly forecasting. The method uses exponential smoothing together with recurrent neural network (RNN). The ES and seasonal parameters are specific to each series, while the neural network weights are shared. Information on the data category are used in the RNN. This method is quite complex, and we try to understand why it performs well through separate experiments. Note that it is slow, especially for long series: hourly forecasts took three days on an 8-core Intel Xeon E5 2667v3 computer — *Cardt* takes a couple of seconds.

The following table is based on 10 000 replications from DGP (10) with default parameters using $T = 28$ with $H = 6$ out-of-sample forecasts. The results from 118 are obtained using the code provided on Github. It shows that 118 retains an advantage for annual data. While there is a concern that any whole database method for M4 will use future data, this cannot happen in the DGP, thus confirming the advantage of 118.

	Mean					Median				
	RMSE	MAPE	MASE	sMAPE	MAAPE	RMSE	MAPE	MASE	sMAPE	MAAPE
118	106.89	13.87	3.00	13.06	12.84	27.10	9.40	2.20	9.40	9.33
Cardt	115.43	13.86	3.24	13.92	13.02	28.76	10.11	2.36	10.36	10.03
THIMA.log	122.05	14.46	3.41	14.43	13.48	29.60	10.69	2.45	10.92	10.61
Theta	131.03	15.45	3.86	16.07	14.48	32.10	12.01	2.70	12.47	11.88
Naive	145.83	16.75	4.46	17.79	15.85	37.70	13.93	3.10	14.64	13.76
118.sep	141.33	18.03	4.29	17.80	16.78	36.80	13.86	3.20	14.11	13.66

	MASE at horizons			
	1	2	4	6
118	1.37	2.06	3.27	4.67
Cardt	1.41	2.15	3.54	5.18
THIMA.log	1.48	2.24	3.73	5.47
Theta	1.53	2.43	4.25	6.36
Naive	1.63	2.71	4.94	7.51
118.sep	2.25	2.92	4.56	6.75

Next, we apply the 118 method to each series in isolation. This is in the table as *118.sep*: now forecast accuracy is on a par with the Naive forecasts.

References

Assimakopoulos, V. and K. Nikolopoulos (2003). The theta model: a decomposition approach to forecasting. *International Journal of Forecasting 16*, 521–530.

Bates, J. M. and C. W. J. Granger (1969). The combination of forecasts. *Operations Research Quarterly 20*, 451–468. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Bergmeir, C. and J. M. Hyndman, R. J. Benítez (2016). Bagging exponential smoothing methods using stl decomposition and Box–Cox transformation. *International Journal of Forecasting 32*, 303–312.

Castle, J. L., J. A. Doornik, and D. F. Hendry (2018). Selecting a model for forecasting. Discussion paper 861, Department of Economics, University of Oxford.

Clements, M. P. and D. F. Hendry (1993). On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting 12*, 617–637. Reprinted in T.C. Mills (ed.), *Economic Forecasting*. Edward Elgar, 1999.

Doornik, J. A. (2013). *Object-Oriented Matrix Programming using Ox* (7th ed.). London: Timberlake Consultants Press.

Doornik, J. A., J. L. Castle, and D. F. Hendry (2019). Card forecasts for M4. *International Journal of Forecasting*. submitted.

Engler, E. and B. Nielsen (2009). The empirical process of autoregressive residuals. *Econometrics Journal 12*, 367–381.

Ermini, L. and D. F. Hendry (2008). Log income versus linear income: An application of the encompassing principle. *Oxford Bulletin of Economics and Statistics 70*, 807–827.

Findley, D. F., B. C. Monsell, W. R. Bell, W. R. Otto, and B.-C. Chen (1998). New capabilities and methods of the X-12-ARIMA seasonal-adjustment program (with discussion). *Journal of Business and Economic Statistics 16*, 127–177.

Gneiting, T. and A. E. Raftery (2007). Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association 102*, 359–378.

Goodwin, P. B. and R. Lawton (1999). On the asymmetry of the symmetric MAPE. *International Journal of Forecasting 15*, 405–408.

Hyndman, R. J. and B. Billah (2003). Unmasking the theta method. *International Journal of Forecasting 19*, 287–290.

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

Hyndman, R. J., A. B. Koehler, J. K. Ord, and R. D. Snyder (2008). *Forecasting with Exponential Smoothing*. New York: Springer.

Hyndman, R. J., A. B. Koehler, R. D. Snyder, and S. Grose (2002). A state space framework for automatic forecasting using exponential smoothing. *International Journal of Forecasting 18*, 439–454.

Hyndman, R. J., M. O'Hara-Wild, C. Bergmeir, S. Razbash, and E. Wang (2017). R package 'forecast', version 8.2. Technical report, CRAN.

Kim, S. and H. Kim (2016). A new metric of absolute percentage error for intermittent demand forecasts. *International Journal of Forecasting 32*, 669–679.

Koehler, A. B. (2001). The asymmetry of the sAPE measure and other comments on the M3-competition. *International Journal of Forecasting 17*, 570–574.

Ladiray, D. and B. Quenneville (2001). *Seasonal Adjustment with the X-11 Method*. Berlin: Springer Verlag.

Legaki, N. Z. and K. Koutsouri (2018). Submission 260 to the M4 competition. Github, National Technical University of Athens.

Makridakis, S. (1993). Accuracy measures: theoretical and practical concerns. *International Journal of Forecasting 9*, 527–529.

Makridakis, S. and M. Hibon (2000). The M3-competition: Results, conclusions and implications. *International Journal of Forecasting 16*, 451–476.

Makridakis, S., E. Spiliotis, and V. Assimakopoulos (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PLoS one 13*, 176–179. doi:10.1371/journal.pone.0194889.

Smyl, S. (2018). Submission 118 to the M4 competition. Github, Uber Technologies.

Spanos, A., D. F. Hendry, and J. J. Reade (2008). Linear vs. log-linear unit root specification: An application of mis-specification encompassing. *Oxford Bulletin of Economics and Statistics 70*, 829–847.