Executive summary

This report is aiming to use data in M3-competetion to explore various models and use them to generate forecasts as well as evaluate the forecasts through different measures and aspects.

It firstly explores different models in the families of regression, exponential smoothing and ARIMA models to find the most suitable models from each family for a certain time series. Among these three methods, ARIMA model turns out to have a more decent performance than other two after comparing and evaluating these three models. The evaluation of the 18-month forecasts that have been generated from these models also shows a comprehensive advantage of the ARIMA model.

Then, in the batch forecast, the MAPA models is identified as the best model for the overall 130 time series. The equally weighted combination strategy is also proven to be effective on improving or at least stabilizing the accuracy. The selection strategy also has a relatively good performance, however, it takes more time than other methods. The evaluation and analysis indicate that the forecast horizon and data category have significant effects on the accuracy. In this process, the MAPA, equal weights combination and self-defined methods are turned out to be the best three models.

Contents

Modelling and Forecast on M3-Competition

Introduction

This paper mainly has two parts. In the first part of this paper, it will manually fit and test different models in order to choose three models with relative high accuracy with in-sample data. Then it will produce and graph the forecasts from those models and compare and evaluate the performance of these three models by using error measures like RMSE, MPE, MASE and MAPE and statistic tests. In the second part, it will implement three automatic models for 130 time series while two strategies will be designed for selecting and combining models. After that, the five methods will be used to generate forecasts which will also be decomposed, evaluated and analysed with different horizons and data categories.

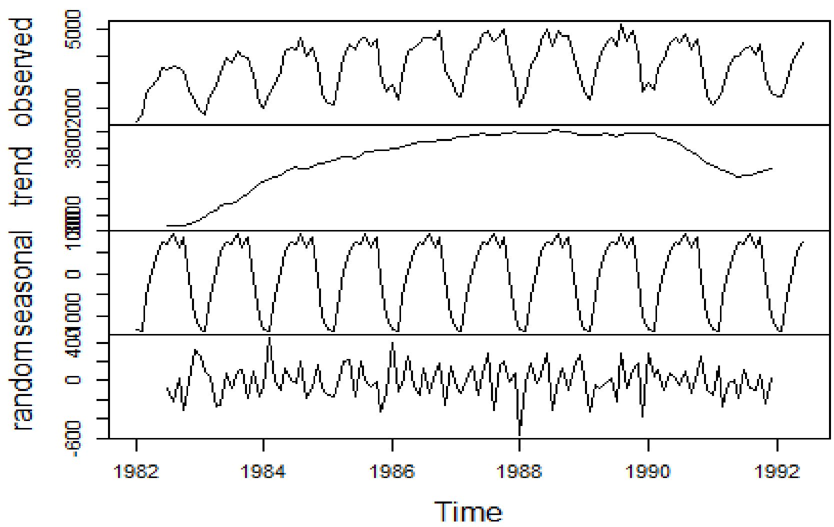
Manual Modelling

In first section of this paper, it will use three models, which are regression model, ARIMA model and exponential model, to produce forecasts for the shipments of cement in Portland which is stored as the 1907th time series in M3-competition. The justification of selecting models and evaluation will be provided for all of these three methods while the forecasts will be compared and commented at the end of this part.

Regression Model

Before choosing the regression model, a decomposition method has been applied by using the R function *decompose()* in order to understand the general characteristics of this data set. Figure 1.1, which is the result of the decomposition, shows it has a strong seasonality while the trend is more piecewise other than monotonously linear.

Figure 1.1: Decomposition of the Time Series



Therefore, a regression model with seasonality and trend was picked as a start model. This was implemented with the R code below,

```
fit.lin1 = tslm(y ~ trend + season)
```

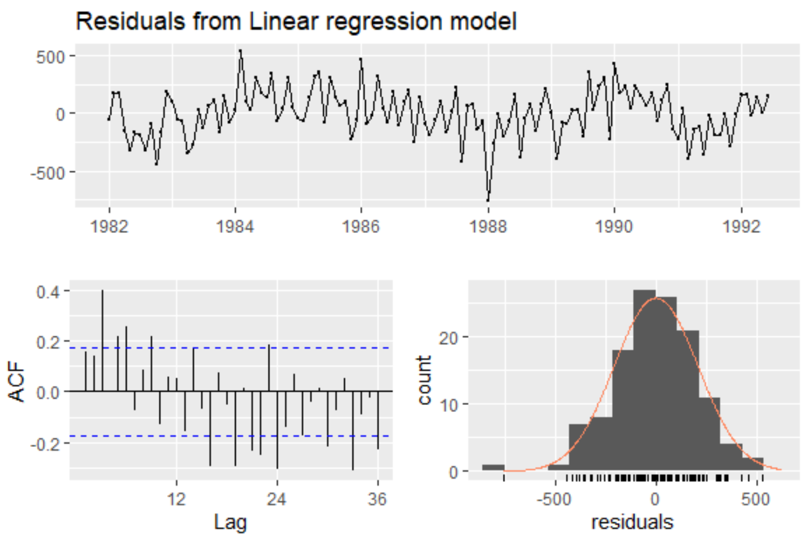
. However, this model did not capture the characteristics thoroughly. From its residuals of *fit.lin1* (as shown in Appendix 1.1), it was known that the residuals from this model is not linear while the ACF plot indicates that it still has a strong correlation. From the trend diagram in Figure1.1, it is evident that during the period from January of 1982 to January of 1988, the shipments of cement in Portland has been increasing almost linearly while after 1988, there was a rapid decrease until the end of 1991. Therefore, two groups of dummy variables have been used to make the model piecewise. According to the changes, January of 1982 and January of 1988 are specified as two knots of the two groups dummy variables as the code below.

```
t <- time(M3[[1907]]$x)
knot1 = 1988 + 1/12       # January of 1988 is the first knot
knot2 = 1991 + 1/12       # January of 1991 is the second knot

dumVar1 = ts(12*pmax(0,t - knot1), start = c(1982,1))
dumVar2 = ts(12*pmax(0,t - knot2), start = c(1982,1))

fit.lin.final = tslm(y ~ trend + season + t + dumVar1 + dumVar2)
```

The whole code will be presented in the accompanying R file.

After adjusting the regression model, the result has been improved to a considerable extent. The adjusted r^2 is increased from 0.846 to 0.950. As for the residuals, as shown in the residuals, as shown in Figure 1.2 below,

Figure 1.2: Residuals of Regression Model

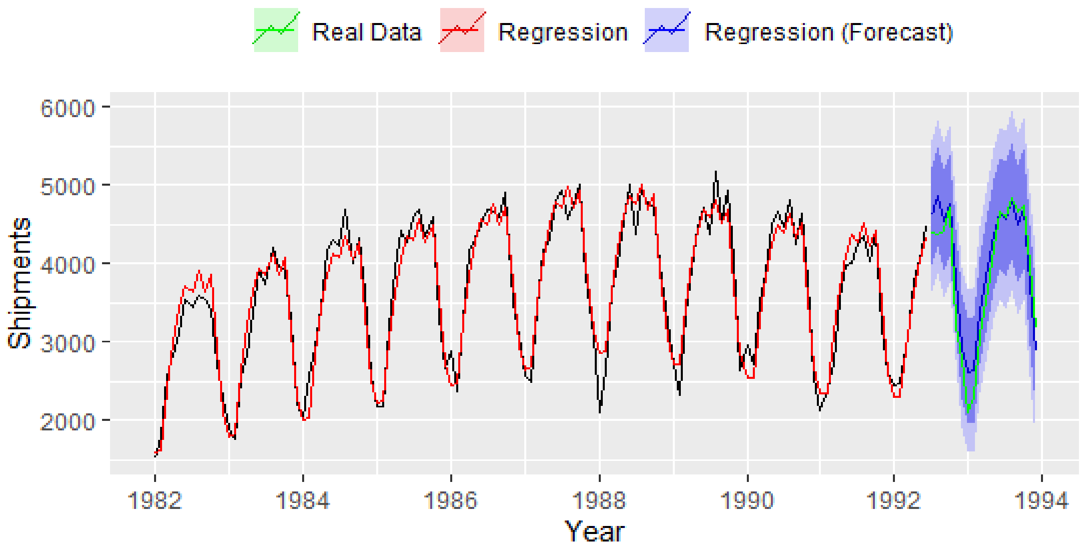


it is evidently not white noises as in the ACF plot there are almost 36.1% of the spikes are lie out of $\frac{\pm 2}{\sqrt{T}}$ (Hyndman and Athanasopoulos, 2018), where T = 126 which represents the length of the time series. Also, from the first plot of Figure 1.2, there are some patterns which indicate that the residuals of this regression model are not totally linear as there are some fluctuations. The p-values of both Box-Pierce test and Ljung-Box test are smaller than 5%

which indicates that there is a part of residuals which can be distinguished from white noises. The remaining autocorrelation may result from sub-optimal dummy variables. Considering further adjusting the variables may lead to overfitting, the two variables that have been used were designed as linear whereas the real trend is not. As for the normality, despite the small distribution in the left, it is close to normal distribution. Also, the p-value of Shapiro-Wilk normality test for this model is 0.5084 which is higher than 5%. Therefore, it can be considered as normally distributed.

Figure 1.3 shows the fitness of the regression model to the in-sample data as well as the forecast the future shipments in the next 18 months. The comment of the performance of this forecast and the model will be presented at the end of this part together with the comparison of other models.

Figure 1.3: The Forecasts of the Regression Model



Exponential Smoothing Model

To choose an appropriate exponential smoothing model, it is necessary to try different combinations. According to Pegel's classification, the current time series is closer to additive trend model. Therefore, a potential start model could be the simple exponential smoothing model, which is represented as "ANN" in the *ets()* function in R.

Starting from the first model, 5 other exponential smoothing models have been tried out. Table 1.1 demonstrates the AICc and three error-measure values of the six exponential smoothing models together with a benchmark model (Seasonal Naïve) for the in-sample data.

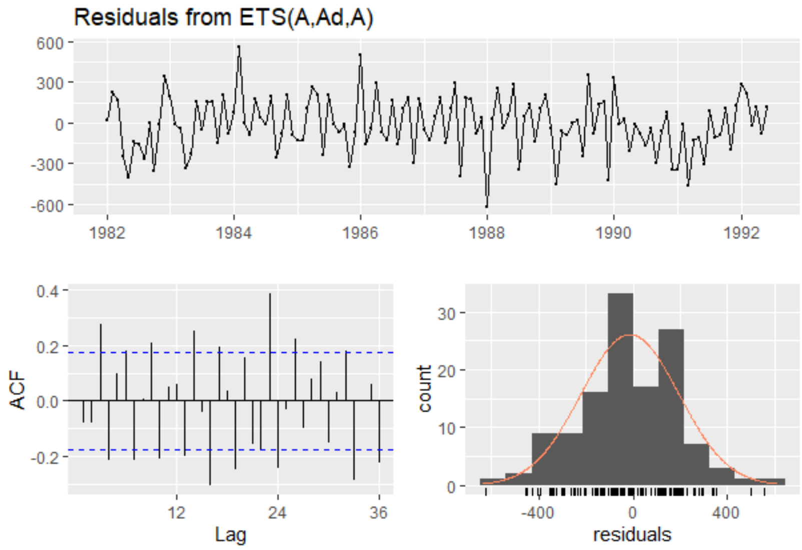
Table 1.1: The AICc and Error Measures of EXP Model

	In-sample Value			
	AICc	**MPE%**	**MAPE%**	**MASE**
Damped Holt-winter EXP Smoothing	1995.371	-0.615	5.009	0.596
Holt-winter EXP Smoothing	1996.910	-0.794	5.192	0.608

Seasonal EXP Smoothing	1998.999	-0.097	5.273	0.618
Seasonal Naïve		1.897	8.101	1.000
Simple EXP Smoothing	2215.839	-0.696	14.229	1.724
Damped EXP Smoothing	2222.224	-1.402	14.385	1.727
Holt EXP Smoothing	2225.293	-3.473	14.656	1.736

According to the MASE and MAPE values, there are three models, which are seasonal, Holt-winter, and damped Holt-winter exponential Smoothing models, having more accurate performance when applied to the in-sample data while they also have a similar AICc values. However, among these three models, the damped Holt-winter exponential smoothing has a more comprehensive advantages compared with the other three as it has the lowest MAPE, MASE and AICc values.

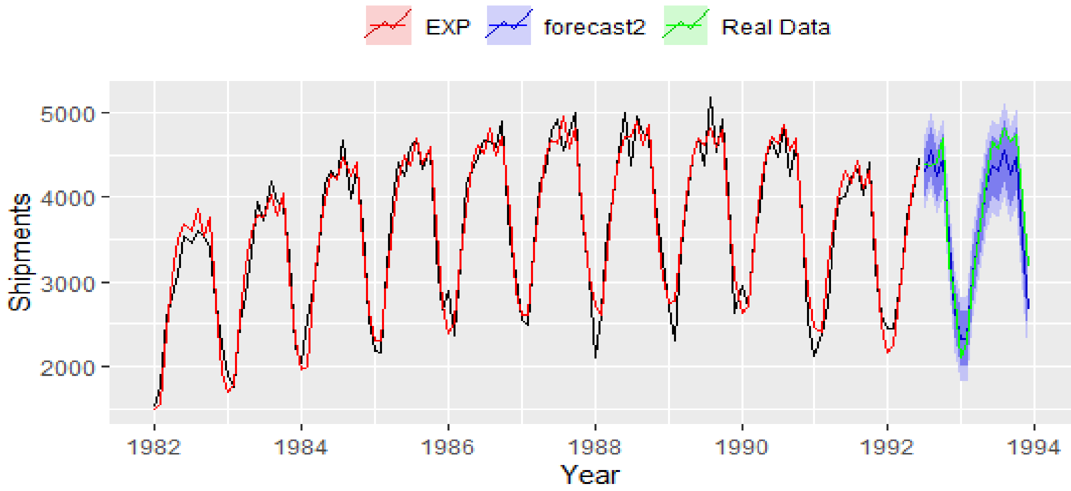
Figure 1.4: Residuals of Holt-winter
Exponential Smoothing Model



After applying the damped Holt-winter exponential smoothing model, the residual graph (Figure 1.4) shows that residuals of this model is more linear than that of the regression model while its normality also seems to be reasonable although it has abnormalities in the middle and p-value of Shapiro-Wilk normality is 61.69% way higher that 5% which indicates the residuals are able to be considered as normally distributed. Whereas, there are also some autocorrelation which are not from white noise. This can be concluded from the ACF plot, which evidently has a certain number of spikes lying out of the limitation and also the factor that both of the Box-Pierce test and Ljung-Box test give p-values that are smaller than 5%.

The forecast of the chosen exponential smoothing model is shown in Figure 1.5.

Figure 1.5: The Forecasts of the Damped Holt-Winter Exponential Smoothing Model



ARIMA Model

In order to apply an appropriate ARIMA model, the first step is to decide how many first-order differencing processes and seasonal differencing processes are needed. To know that, it used Augmented Dickey-Fuller test to see whether the time series is stationary. By running the R function adf.test() with K (the lag order) equals to 12, we see the p-value given is 0.9283 which is a quite high number. This indicates the current time series is non-stationary. Also, the Kwiatkowski-Phillips-Schmidt-Shin give a p-value of 0.0284 which supports the factor that the data set is non-stationary.

To solve the non-stationarity issue, the differencing method should be used. Before that, in order to determine how many times of differencing are needed, two R functions, *ndiffs()* and *nsdiffs()*, were applied. As both of these two functions delivered a value of 1, it was decided to have 1 differencing for trend and seasonality.

After applying two differences, another R function, *tsdisplay()* was used to draw the ACF and PACF plots. Figure 1.6 shows that, in ACF plot, there are 1 significant seasonal spike and 5 significant non-seasonal spikes in the beginning which means there is likely to have a moving average model with q = 5 and Q = 1. Based on this, a possible start ARIMA model could be $ARIMA(0,1,5)(0,1,1)_{12}$.

Meanwhile, the PACF, where there are none significant seasonal spike and 4 significant non-seasonal spikes in the beginning and that means a potential autoregression model with p = 4 and P = 0 may exist. That leads to a possible ARIMA model, $\mathrm{ARIMA}(4,1,0)(0,1,0)_{12}$

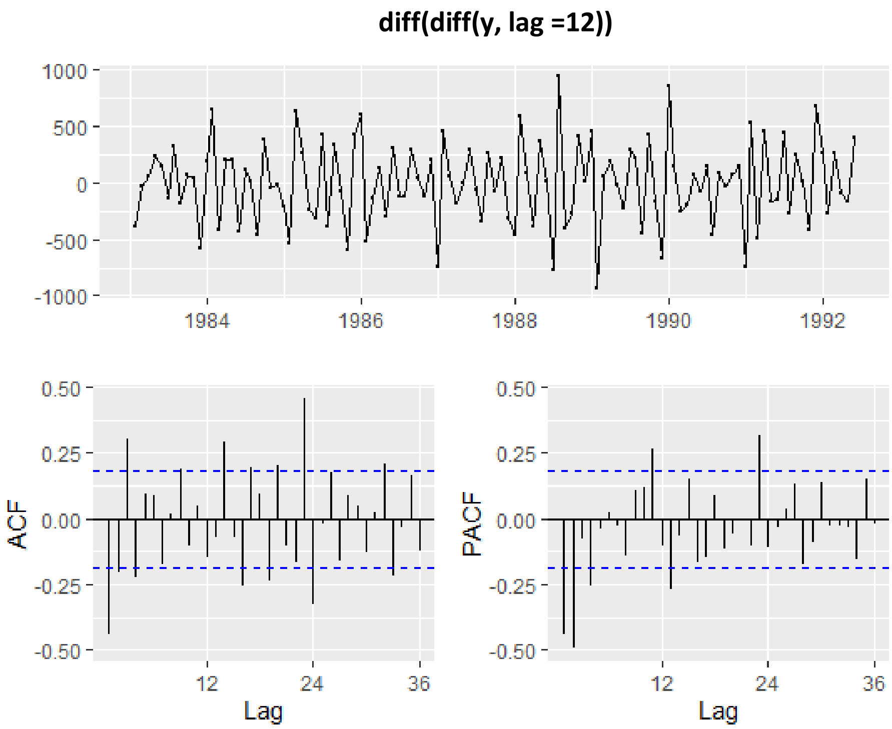


Figure 1.6: ACF and PACF Plots after two Differencing

Start from the two ARIMA models, 27 models (including 12 autoregression models and 15 moving averaging models) in total have been tested. Eight most fitted models have been listed in Table 1.2. As we can see, the moving averaging model with q = 5 and Q = 2 has the lowest AICc, MASE and MAPE values. Thus, the model (ARIMA$(0,1,5)(0,1,2)_{12}$ was chosen in this case. Large q in ARIMA model sometimes may lead to overfitting, whereas in this case, q = 5 was tested to be still appropriate for the data set.

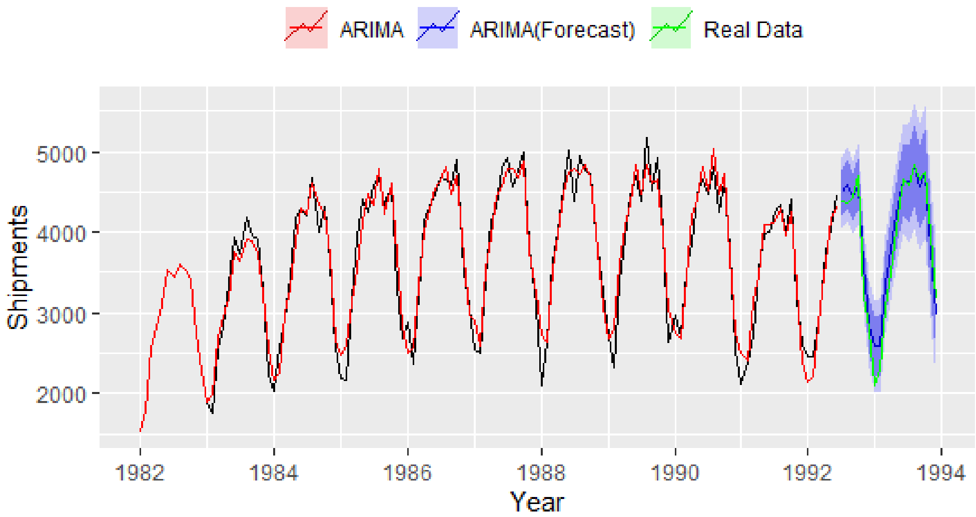
Table 1.2: The Performance of 8 Most Fitted ARIMA Models

ARIMA Models	MPE	MAPE	MASE	AICc
ARIMA(0,1,5)(0,1,2)	-0.990	4.562	0.540	1566.603
ARIMA(0,1,4)(0,1,2)	-1.102	4.571	0.540	1567.265
ARIMA(0,1,3)(0,1,2)	-0.956	4.626	0.549	1568.047
ARIMA(0,1,1)(0,1,2)	-1.248	4.814	0.564	1570.969
ARIMA(0,1,2)(0,1,2)	-1.210	4.782	0.558	1572.115
ARIMA(4,1,0)(0,1,0)	-0.935	4.860	0.574	1573.399
ARIMA(2,1,0)(0,1,0)	-0.700	5.263	0.621	1577.391
ARIMA(1,1,0)(0,1,0)	-0.565	7.012	0.837	1628.073

Through the residual plots (see in appendix 1.2) of the chosen ARIMA model, it can be seen that the residuals of the ARIMA model is near-linear while there are still around 13.89% (5 out of 36) of the spikes are significant. The p-value of Box-Pierce and Ljung-Box tests are 0.666 and 0.597 respectively, which shows a great process when compared with the previous two models. Also, the Shapiro-Wilk normality test gives 0.056 as its p-value which means the residuals can be considered as normally distributed.

The forecast of the chosen ARIMA model is presented in Figure 1.7.

Figure 1.7: The Forecasts of The ARIMA Model $\mathbf{ARIMA(0,1,5)(0,1,2)_{12}}$



Comments and Comparison

From the three forecast plots (Figure 1.3, Figure 1.5, Figure 1.7), it can be roughly concluded as that, the ARIMA model has the most accurate forecast with the horizon of 18 months while the regression model and exponential model both have some insufficient in forecasting. This is likely due to the autocorrelation that exists in the residuals of these two models.

Table 1.3 shows the forecasting performance of the above three chosen models. From this table, it is evident that all of these three models have a more accurate forecast than the benchmark model (seasonal naïve) as they have smaller RMSE, MPE, MAPE and MASE than the naïve method while the ARIMA model is obviously the most suitable for this data set.

Table 1.3: The Performance of three Chosen Models and one Benchmark Model

	Out-sample Values			
	RMSE	**MPE**	**MAPE**	**MASE**
ARIMA	206.659	0.218	5.075	0.647
Damped Holt-winter EXP Smoothing	259.080	2.805	5.714	0.794
Regression	279.923	-4.678	7.004	0.820
sNaive	359.753	4.741	7.649	1.056

The scale-dependent error RMSE indicates that when focus on the forecast of mean, the ARIMA model will be the most appropriate one. MPE values shows that expect the regression model, all other three model have over-forecasted. Refer to the MAPE values, we are able to know that both ARIMA and exponential models has averagely less than 6% error while the average error percentage of the other two models exceed 7%. The MASE, which in this case should be defined by the seasonal naïve method, is the straightest measure to compare these four models. According to the formula of MASE (Hyndman and Athanasopoulos, 2018), it is easy to calculate that the ARIMA model has only around 61% error that the seasonal naïve method has while the exponential and regression models have 75% and 78% respectively.

Batch Forecasting

In this part, three automatic models as well as two model strategies will be introduced to produce the forecasts across 130 time series which were collected from 6 fields. The process and reason of constructing the two strategies will be justified in detail. These five selections will then be evaluated with three error measures. Further, it will discuss the performance of these five methods under different situations.

Data and Models

130 time series have been chosen between M3[[1500]] and M3[[2800]] with the condition that all of them should end with digital "6". All of the 130 time series are monthly data from various fields. Thus, the frequencies for all the data sets in this project are assumed as 12.

The three automatic models that will be used in this project are automatic exponential smoothing, automatic ARIMA and multiple aggregation prediction algorithm (MAPA) models. As for the two strategies, the first one is a model selection strategy which uses cross-validation to compare accuracy among the above three automatic models while the second strategy combines the forecasts of all four methods that have been applied in order to improve the accuracy for the forecast.

Three Automatic Models

The MAPA model was implemented by using the function *mapa()* from the package *MAPA* in R while the automatic ARIMA and exponential smoothing models are respectively applied with auto.arima() and *ets()* functions from package *forecast*. The relevant code is included in accompanying R files.

The Model Selection Strategy

This strategy consists of two sections, the validation and forecast. In the validation part, it will check errors of the three candidate models for every data sets. After comparison, it then records the best model in a list for each time series. In the forecast part, according to the best model list, the most suitable model will be chosen to produce the forecast for every time series.

In most of the real cases, the future observations are not always available when fitting the suitable models. Therefore, for the purpose of stimulating the reality, we assume the testing (out-sample) data is not accessible for the selection part. In order to choose the most appropriate model for each time series, the cross validation has been used to evaluate the three automatic models.

The Cross-Validation

In this procedure, it was decided to have an initial window length of

$$W = Length(Y) - 2 * H + 1$$

where W is the window length, Y is the original data set and H is the forecast horizon. Under this arrangement, the data for validating will have exactly the same length of the forecast horizon. Moreover, for every validation, it will have three steps. This is because the cross-validation usually takes a huge amount of time. Considering that we have three models to test, if each data set needs 12 steps of validation then validating one series requires to fit 36 models. This could be considerable expensive for 130 time series. In fact, it needs almost 3 hours to only process the cross validation under the computer with Windows 10 system and Intel Core i5-8265U, 8GB RAM. By adjusting the number of validation steps to three, the process time reduced to averagely 33 minutes.

During the validation, in order to choose the most suitable model, it uses the error measure MAE to evaluate the performance of the three candidate models. This is because MAE is easy to calculate and understand, and the comparison in cross validation is only horizon. By selecting the model with the smallest MAE, it means that the fitted values of the chosen model are closer to the read in-sample data which indicates it is more likely to capture the characteristics. Thus, it has the larger chance to produce an accurate forecast.

The Combination Strategies

Combination Strategies are highly praised and advised from a half century ago because it can evidently improve the accuracy of forecasts by, as Kenneth and Robert claimed (Lichtendahl and Winkler, 2020), increasing the diversity of forecast methods, reducing the accuracy risk and improving the robustness. In this section, it will introduce two combination strategies with two different groups of weights.

Although, for achieving a more accurate result, it is recommended to use 5 to 10 forecasts (Gaba, Tsetlin and Winkler, 2017). Considering the main purpose of this part is comparing the three automatic models as well as the selection strategy, it will combine the only four forecasts that have been generated from the earlier procedures.

The Equally Weighted Strategy

According to J. Scott Armstrong, when there is no evident clue showing that an unequal weighting of forecast should be use, it is suggested to use equal weights for combing forecasts (Armstrong, 2001). In this case, without further researches, it is not likely to find any significant evidence for unequal weighting. Therefore, a group of equal weights is chosen as a start point and also the first combination strategy under which all these four forecasts have a weight 0.25.

The Unequally Weighted Strategy

The second unequally weighted strategy was designed based on the previous one. Considering a part of models may over forecast the result while the others under forecast, the idea of the unequally weighted strategy is to find the balance between over-forecast models and under-forecast model. To realize this strategy, the error measure MPE has been introduced since it can reflect both the positive and negative errors. The MPE values have been calculated from all the four methods with the in-sample data of 130 time series. Therefore, for every time series, it will have four MPEs.

These MPEs can be all positive or negative, one positive/negative and three negative/positive, as well as two positive and two negative. Under the first situation, it is no possible to find the balance anymore. Hence, it will be the same with the equally weighted strategy. For the second situation, it will firstly calculate the mean of the three positive/negative MPEs and give equal weights to those three forecasts. Then it uses the mean MPE to compare with the negative/positive MPE and find a weight that can make the overall MPE zero. For example, there are 4 MPEs, A, B, C, D with first three are positive and the last negative. The mean MPE of the first three is $P = A \times \frac{1}{3} + B \times \frac{1}{3} + C \times \frac{1}{3}$. Then, to let overall MAP become zero the weights for P and D should be subject to,

$$P \times \gamma_p = -D \times \gamma_D,$$

and

$$\gamma_p + \gamma_D = 1,$$

which leads to,

$$\gamma_p = \frac{D}{P - D},$$

$$\gamma_D = \frac{P}{P - D}.$$

Thus, the weights for A, B and C are $\frac{\gamma_p}{3}$.

In the last situation, for instance A, B, C and D are subject to the following conditions,

$$A > B > 0 > D > C,$$

$$\gamma_A + \gamma_B + \gamma_C + \gamma_D = 1,$$

the following weights should make the overall MPE zero,

$$\gamma_A = \frac{D}{A + B + C + D},$$

$$\gamma_B = \frac{C}{A + B + C + D},$$

$$\gamma_C = \frac{A}{A + B + C + D},$$

$$\gamma_D = \frac{B}{A + B + C + D}.$$

Error Measures

In this part, three error measures have been used to evaluate the performance of the forecasts. The first one is MAE as it easy to understand, compute as well as interpret. This measure will be used to evaluate the five methods horizontally since it is scale-dependent and cannot be used to compare between different time series (Hyndman and Athanasopoulos, 2018).

The second measure is sMAPE which is also easy to calculate. The reason for choosing it is that it is scale-independent which means it can be used in the vertical comparison. Meanwhile, it is also symmetric which can avoid giving more penalty on the negative errors like MAPE does (Hyndman and Athanasopoulos, 2018).

The final error measure is MASE. As an alternative for comparing across time series (Hyndman and Athanasopoulos, 2018), can be used either horizontally or vertically when comparing with the benchmark method that is seasonal naïve model.

The Analysis and Evaluation

After integrating data and calculation, the performance of each methods is presented as three error measure values in Table 2.1. From the table, we can see that according to the MASE and sMAPE value, MAPA model is the most accurate method for produce 12 months forecasts of the 130 time series. The equally weighted combination strategy is very close to the first position in both MASE and sMAPE while the unequally weighted combination, on the opposite, has a considerably poor performance in forecasting. This is very likely because of the inappropriate weight setting.

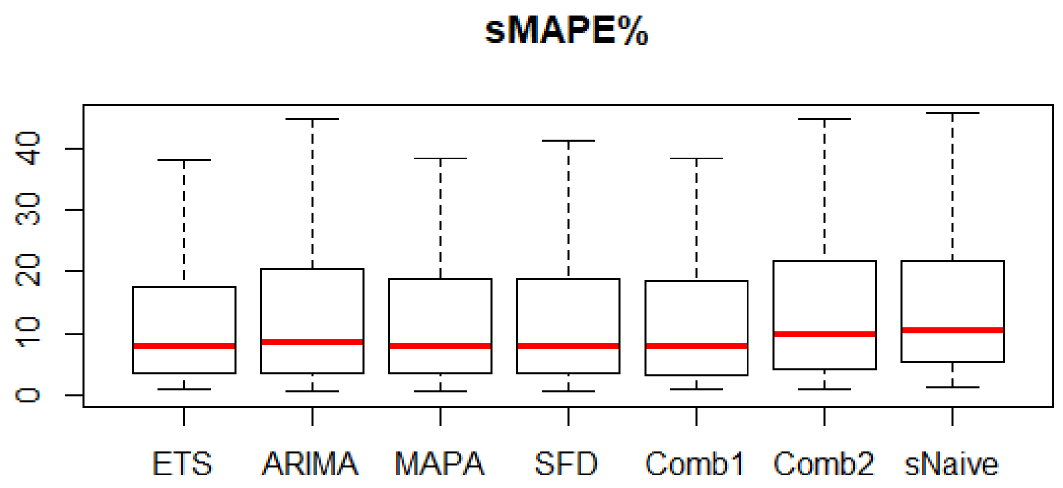
Table 2.1: The Performance of 9 Methods with Horizon of 12 Month

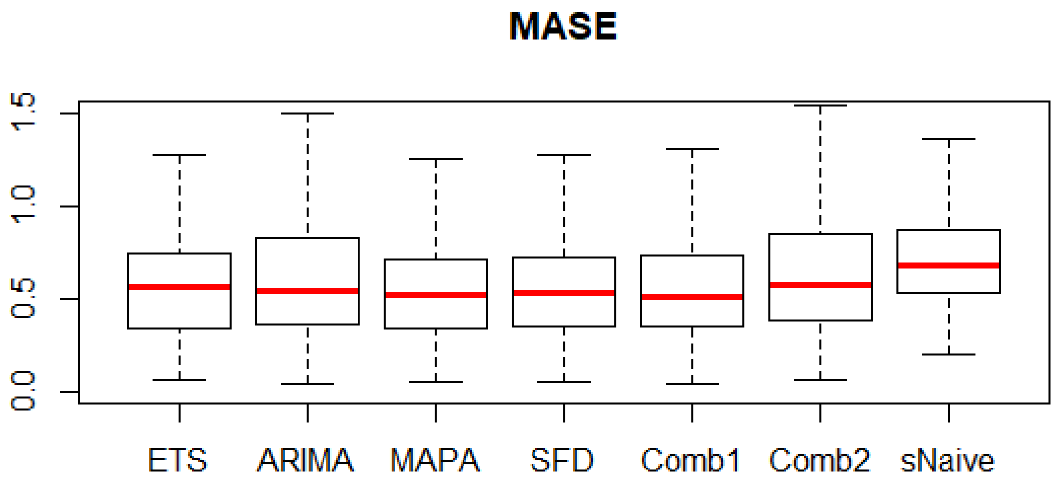
Models	**MAEs**	**MASE**	**sMAPE**
MAPA	510.2331	0.6274	12.5006
Equal Weights Combination	507.2365	0.6375	12.5776
Self-Defined Strategy	517.6165	0.6571	12.7957
Damped Exponential Smoothing-Benckmark	525.3994	0.6583	12.6715
ETS	515.8659	0.6695	12.8231
ARIMA	543.1136	0.7047	13.8884
sNaive-Benckmark	675.6489	0.9296	15.2116
Unequal Weights Combination	671.5382	1.0262	15.4541
Naïve- Benckmark	735.2446	1.1524	17.517

The self-defined method, which is in the third position, is also close to the previous two. Compared with the three benchmark methods, the self-defined method has an advantage on the MASE and MAE measures while the sMAPE value of the damped exponential smoothing model is slightly lower. Consequently, it can be concluded as that the self-defined method has an accuracy advantages against the three benchmark models while the performance of the damped exponential model is quite close to that of it.

Additionally, from Figure 2.1, we can also see more straightforward from the sMAPE boxplot and MASE boxplot that the MAPA, SFD (self-defined) and Comb1(equally weighted combination strategy) methods have less and smaller outliers than the rests. The comparison between ARIMA model and the weighted combination strategy indicates that combination is effective to improve the accuracy or at least it makes the forecasts more stable, as the large outliers of ARIMA are neutralized to some extent.

Figure 2.1: The Error Measures for 12 Months Forecasts of The Six Methods Together with The Benchmark Method across 130 Time Series



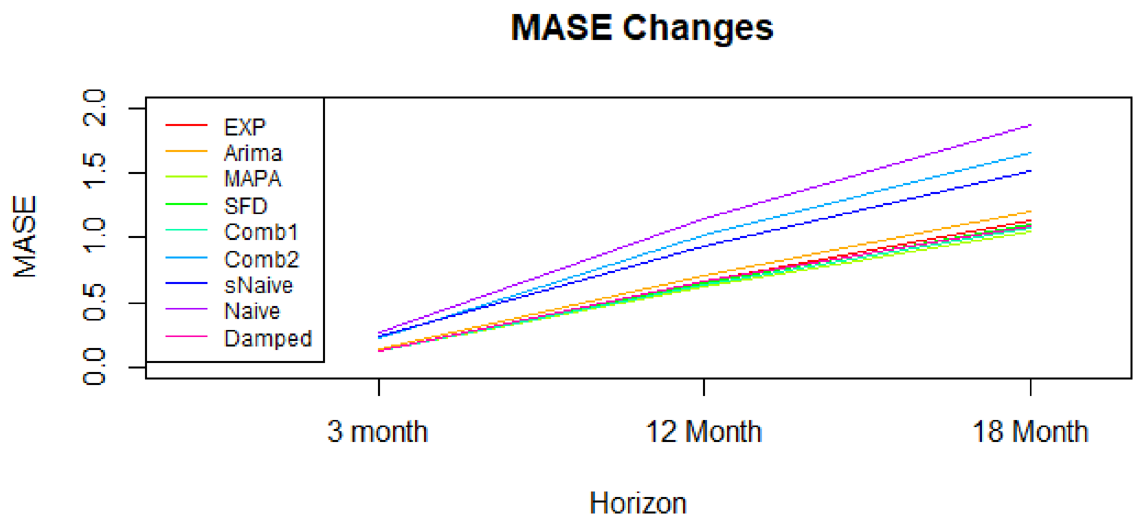


Analysis on the Effect of Different Horizon

In order to evaluate these methods more comprehensively, the changes in the mean of total MASE according to different length of horizons have also been calculated for each method. In this paper, it mainly classified three horizons, 3 months (short-term), 12 months (medium-term) and 18 months (long-term).

From Figure 2.2 we can see that, by adding the forecast horizon, the forecast accuracy of all the methods are dropping down even though with different rates. That means, with the expanding of forecast horizon, the difference of performance between these methods are also growing. The top three methods that has been mentioned above also have a growing trend in errors, whereas, the gap between these three are not obvious even in the long-term horizon while the difference between the top three and the two naïve models are expanding rapidly. Meanwhile, the ARIMA model stays quite close to the top class now but it can be foreseen that with a longer horizon, the forecast of it will also differ to a large extent.

Figure 2.2: The Change of MASE in Different Horizons



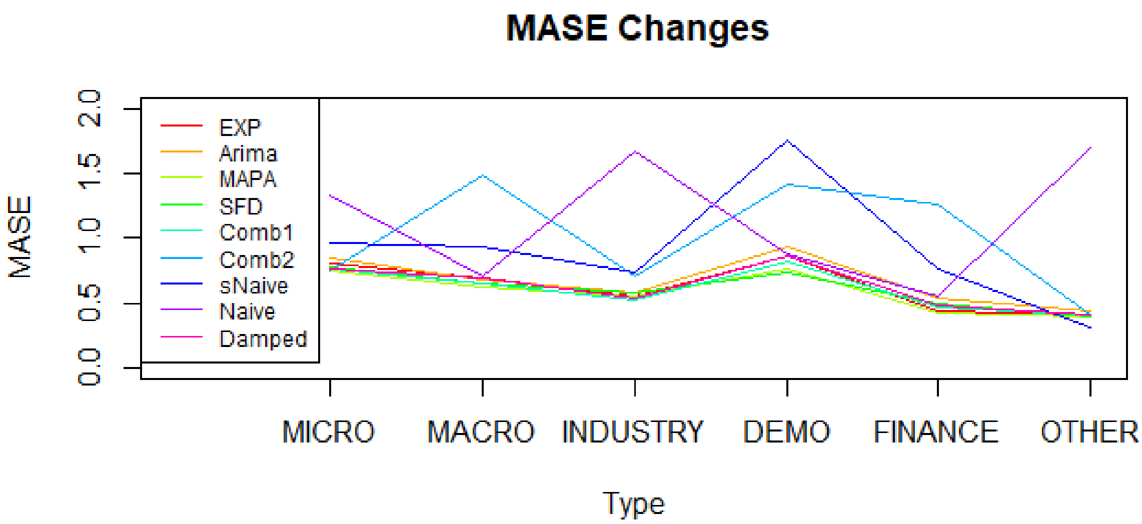
In conclusion, among these 9 methods, the Comb2, sNaive and Naïve models are not suitable for long term forecasting while ARIMA model may also become poorer when the horizon increases.

Analysis on the Effect of Different Data Categories

Expect from the length of forecast horizon, different type of data sometimes also will make a difference to the accuracy of forecasts. All the data sets in this paper can be summarized as 6 categories, which are micro, macro, industry, demographic, finance and others. In order to analysis the performance of these methods vertically, the forecasts data that has been produced were categorized accordingly into six groups. Figure 2.3 shows the changes of MASE value when referred to different data types.

From the figure we can seen that expect from the two naïve methods and the unequally weighted combination, all other methods have a similar performance across different data types. Generally, all these models have better performance in industry, finance and other data while micro data and demographic data are the most difficult to predict. This may because the industry and finance fields are more regular and more consistent to the market.

Figure 2.3: The Change of MASE with Different Data Sources



Conclusions

In summary, this report consists of two main parts. The first part has introduced the process of fitting and choosing the three models for a certain time series according to its characteristics. Then, it has generated and graphed the forecasts by using these models. The evaluated and comparison of appropriateness between the three chosen models is based on the three error measures and several statistic tests. In the second part of the paper, a batch forecasting for 130 time series has been implemented with three automatic models and two strategies of selecting and combing models. The selecting strategy used cross-validation to assess the fitness of the automatic models to the in-sample data in order to choose the most suitable one for each time series. In the combination strategy, it has implemented two methods with equal and unequal weights. It turned out in the evaluation, the unequal weights combination method has a poorer performance than the equal weights combination, which may result from inappropriate weights setting. The evaluation and analysis indicate that with the expanding of the forecast horizon, the accuracy is supposed to decrease and poor models may be affected more by the horizon. Moreover, the data category is also a factor that will influence the accuracy of forecasts. In this paper, the industry, finance data are relatively easier for forecasting while the micro and demographic data could more challenging.

In the process of fitting models in manual section, there was a part of autocorrelation that can not be extracted. This directly affected the fitness of the models. In the future, it may need to add more parameters to cope with the autocorrelation. Additionally, in the process of designing the combination strategy, it was too ideal which leads to the poor performance of the unequally weighted strategy. It needs more scientific supports in order to achieve a suitable group of unequal weights.

References

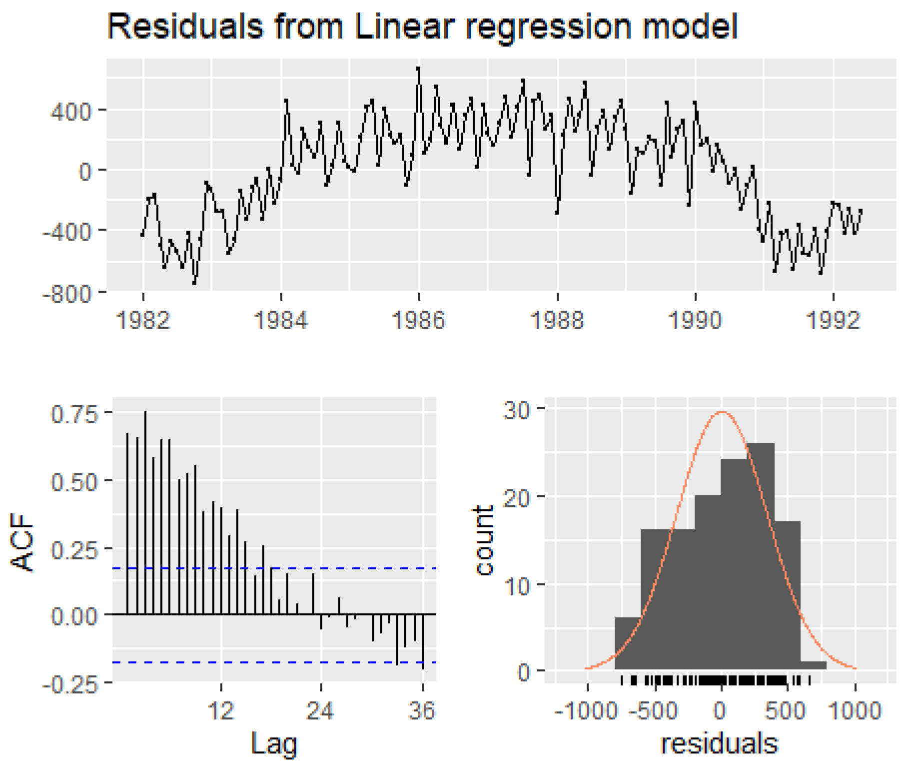
Hyndman, R. and Athanasopoulos, G. (2018). *Forecasting: Principles and Practice*. 2nd ed.

Lichtendahl, K. and Winkler, R. (2020). Why do some combinations perform better than others?. *International Journal of Forecasting*, 36(1), pp.142-149.

Gaba, A., Tsetlin, I. and Winkler, R. (2017). Combining Interval Forecasts. *SSRN Electronic Journal*.

Armstrong, J. (2001). *Principles of Forecasting: A Handbook for Researchers and Practitioners*. Springer.

Appendix 1.1: The Residuals of Start Regression model



Appendix 1.2: The Residuals of ARIMA(0,1,5)(0,1,2)