T-61.5910 – Research Project in Computer and Information Science

A Comparison of Univariate and Multivariate Methods to Forecast the Economy

Advisor:

Pyry Takala

Report by:

Rodrigues Pereira, Ronnie – 522711

2015-12-27

Contents

1 INTRODUCTION

Financial indicators summarize the state of the economy, thus it consists of timed records of its volatility. Furthermore, the forecast of these indexes pose an advantageous opportunity when trading in the stock market. This report presents univariate and multivariate models to forecast the economy. Furthermore, it evaluates three strategies to test the limits of the forecasting window. In other words, the number of days before systematic failure of the predicted forecasts. The evaluated univariate models are Autoregressive (AR), Moving Average (MA), and Autoregressive Integrated Moving Average (ARIMA) while the multivariate are Vector Autoregressive (VAR) and Random Forest (RF).

This report consists of five sections. The first briefly describes the theory of each model as well as the concept of stationarity. The following section presents the experimental design, the assumptions on each approach, and a validation scheme based on simulated data. The third section contains the results of the analysis on the artificial data and on the financial indicators. The next section presents a detailed discussion of these results divided by data set. Furthermore, it explores comparisons between forecasting strategies and models. Lastly, the final section summarizes the findings of this study.

1.1 Data Structure

The financial indicators are variables that evolve with time. In this sense, it belongs to the class of time series data. Furthermore, the univariate models require a stationary time series. Therefore, the following paragraphs describe this type of data as well as the requirements for stationarity.

1.1.1 Stationarity

Linear forecasting models often require stationary data. In other words, the prediction of out of sample results depends on a process that shows no trend and fluctuates around a constant value independent of time. The most stringent definition of stationarity states that the joint distribution of a subset of an indicator **Z** (z_1, z_2, …, z_k) is equivalent to any other subset of **Z**. In contrast, a data set is weakly stationarity if it satisfies two conditions: i) constant mean, regardless of subset; and, ii) covariance dependent only on the distance between the data points (Tsay, 2010). Moreover, NIST/SEMATECH (2012) argue that weak stationarity should also include constant variance and no periodical trends.

Figure 1.1 shows an example of a time series that requires preprocessing before forecasting. This process constitutes the deconstruction of the data into its stationary form. In this sense, an approximately linear trend is subtracted from the data. The resulting variable still shows a pattern of behavior (seasonal). As a result, it requires data transformation to reach stationary (random process).

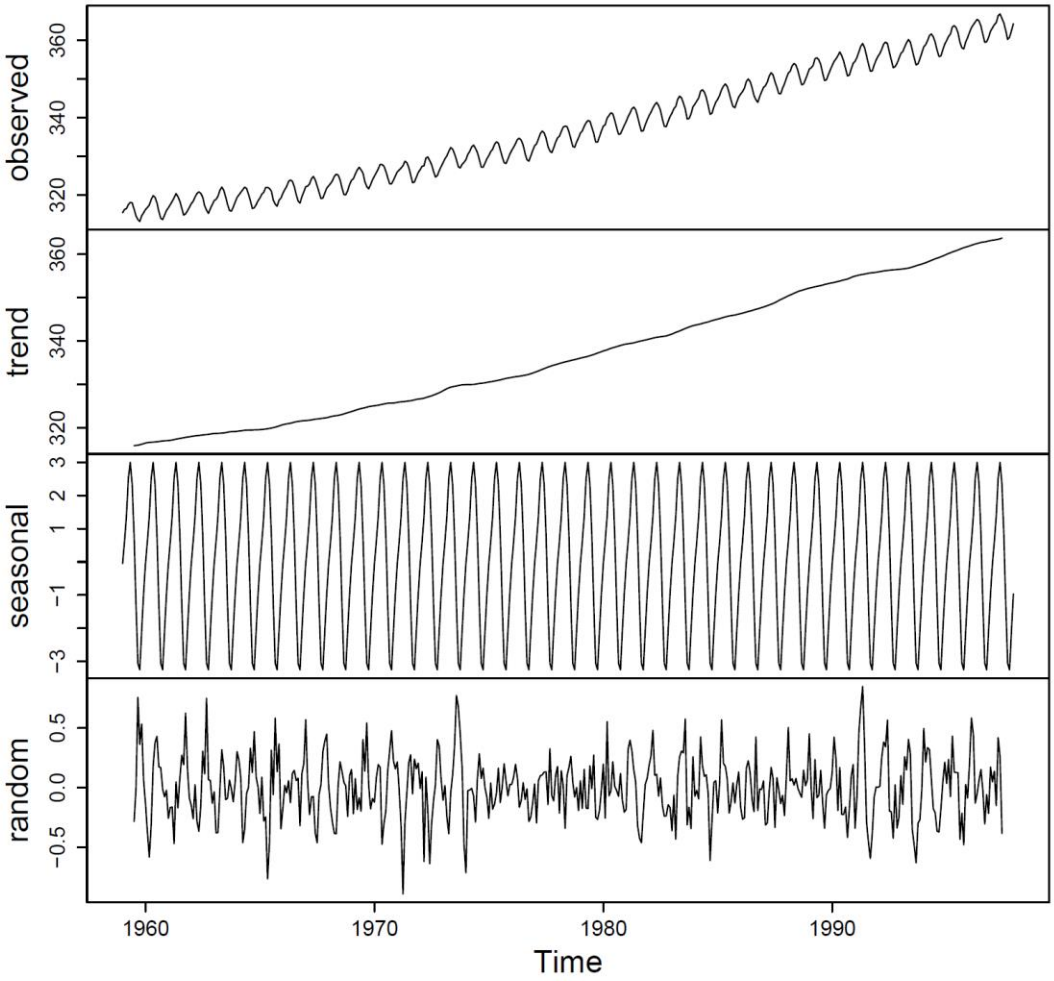

Figure 1.1 – Decomposing a time series. (Bontempi, 2013)

This transformation usually involves simple mathematical manipulation. In other words, positive data may reach stationarity by rescaling operations, such as logarithmic or the by applying the square root. In contrast, negative data may depend on its absolute value or on the addition of a constant (NIST/SEMATECH, 2012). However, the most general approach is to lag the time series, that is, construct a new time series (**Y**) based on differences of the past values of the original data (**Z**) (Bontempi, 2013).

$$Y_i = Z_i - Z_{i-1}$$

Equation 1

$$Y_i = (Z_i - Z_{i-1}) - (Z_{i-1} - Z_{i-2})$$

$$Y_i = (Z_i - Z_{i-1}) - (Z_{i-1} - Z_{i-2}) - (Z_{i-2} - Z_{i-3})$$

Equation 2

$$Y_i = Z_i + \sum_{j=1}^{k} \beta_j \, Z_{i-j}$$

Equation 1 depicts the simplest case that consists of a time series that depends only on its previous value. In contrast, Equation 2 presents the generalization for *k* lagged terms (Nau, 2015).

1.1.2 Augmented Dickey Fuller Stationarity (ADF) Test

This test statistic evaluates the presence of a unit root in the coefficients of an autoregressive time series. Therefore, the null hypothesis of the test denotes the presence of a unit root and its rejection depends on the level of confidence of the analysis. This corresponds to evaluate the ADF test statistic on a selected number of differences, in order to determine the data suitability to forecasts. As a result, the test serves two purposes, that is, to determine stationarity as well as the number of lags to reach this state (Ng and Perron, 1995).

$$\Delta Y_i = \sigma Z_i + \sum_{j=1}^{k} \beta_j Z_{i-j} \qquad \textbf{Equation 3}$$

Equation 3 depicts the differencing operator applied to a time series **Z** for *k* lags. Furthermore, this data does not show trends. The test evaluates the influence of the term σ, that is, only a negative coefficient is able to refute the null hypothesis. In essence, this equation depicts the most general process of a time series

1.2 Time Series Methods

The ARIMA models describe the time series based on their lags and/or shocks to their differences. Furthermore, the out of sample prediction consists of the minimum linear estimate of the next *h* periods. In this sense, the next section explores the simple and mixed models specifically designed to explain time series data.

1.2.1 Autoregressive (AR)

The AR model describes an observation as a combination of the past *p* values of the time series. In this sense, Equation 4 illustrates this behavior. Furthermore, the term e_i denotes white noise to account for data variability (Tsay, 2010).

$$X_i = \phi_0 + \sum_{j=1}^{p} \phi_j X_{i-j} + e_i \qquad \textbf{Equation 4}$$

This equation has similar characteristics to a linear regression model, thus an estimation of the ϕ_i coefficients depend on the number of lags. Moreover, this equation fully describes the case of univariate models. However, this process extends to the multivariate scenario. As a result, the time series depends not only on time but also on the lags across variables (PennState, 2015).

The evaluation of the dependence of each lag on the outcome renders a threshold for the number of differences. This determination consists of the analysis of the autocorrelation and partial autocorrelation functions (Tsay, 2010).

$$\rho_l = \frac{Cov(X_i, X_{i-l})}{Var(X_i)}$$ **Equation 5**

Equation 5 defines the autocorrelation coefficient of observation l (ρ_l). It is worth mentioning that due to the weak stationarity assumption this coefficient is only dependent on the distance between the observations (l). However, the individual arguments consist of the contribution of p lags, as given on Equation 4. Moreover, Figure 1.2 illustrates the plot of the autocorrelation function. This graph shows a well-contained time series, because the initial lags correspond to the highest decrease in the autocorrelation function.

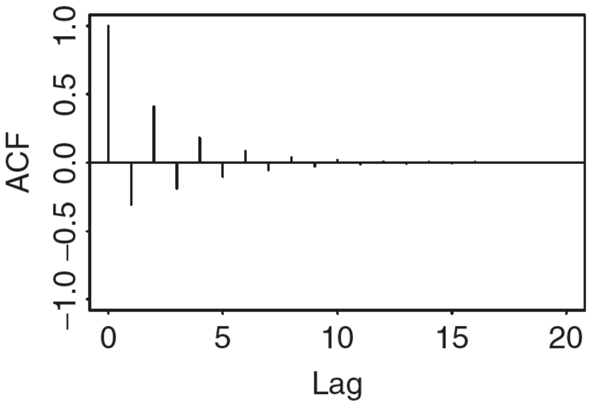

Figure 1.2 – Autocorrelation Function of an AR(2) process. (Tsay, 2010)

The partial ACF differs from the autocorrelation function on the evaluated lags. Its coefficients only include the lag immediately before the observation. In this sense, it demonstrates the added contribution of subsequent lags. In addition, the plot on Figure 1.3 illustrates this behavior. It includes a dotted lines to denote the confidence interval of the given time series.

The number of lags (p) of the AR process should yield a PACF that lies inside the confidence interval. Furthermore, one should be aware of overdifferencing. In this sense, the number of lags consist of the first argument that minimizes the PACF inside of its trust region.

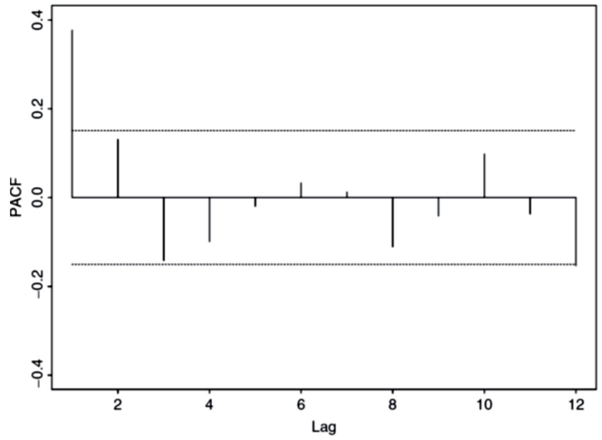

Figure 1.3 – Partial Autocorrelation Function of the US GNP from 1947 to 1991. (Tsay, 2010)

Lastly, the out of sample forecast follows the linear minimum squared error methodology. Hence, it relies on the size of the training sample to act as prior knowledge for the minimization of the prediction error.

1.2.2 Vector Autoregressive (VAR)

This approach extends the concept of the autoregressive model to a multivariate problem. Consequently, Equation 4 should also include coefficients that encompass lagged terms across variables (PennState, 2015). Furthermore, the number of lags depends on the generalization error of the validation set.

1.2.3 Moving Average (MA)

This model is similar to AR in that it relies on past observations, however, it differs on the identity of these data points. The MA model consists of describing an observation as the additive behavior of shocks to the mean. In other words, e is a white noise error term and θ are the parameters of the model (Tsay, 2010), as described on Equation 7.

$$X_i = \mu + e_i + \sum_{j=1}^{q} \theta_j \, e_{i-j}$$

Equation 6

Moreover, the autocorrelation function gives the number of shocks (q) required to describe the data by the MA model. Thus, q is equal to the first argument that minimizes the autocorrelation function inside its confidence interval.

1.2.4 Mixed Models (ARMA/ARIMA)

The combination of the simple AR and MA models give the mixed model ARMA. This is a compromise between the two simpler models and its performance is preferred over the previous only if the former leads to large generalization error. This discrimination is due to the parsimony criteria that relies on the utilization of models with as few parameters as possible (Coghlan, 2015). In essence, the ARMA (p, q) model consists of a concatenation of Equation 4 and Equation 6 (Tsay, 2010).

$$X_i = C + e_i + \sum_{j=1}^{p} \phi_j X_{i-j} + \sum_{j=1}^{q} \theta_j e_{i-j} \qquad \textbf{Equation 7}$$

Moreover, the ARIMA model follows the same behavior outlined on Equation 7. However, it relies on a prior differencing of the input, in order to reach stationarity. Thus, the integrated term stands for the number of differencing operations (d) on the ARIMA (p, d, q) process.

1.3 Machine Learning

The second class of models does not rely on the time series behavior of the data set. In fact, the implementation of the random forest algorithm is common in machine learning applications. Hence, this study briefly describes the concept of this methodology.

1.3.1 Random Forest

This ensemble technique averages the result of many decision trees. Therefore, the training data is important, in order to extend the trees with as many decision rules as possible. In other words, the rules group similar data points together. These groups are called leafs, in this sense, pure leafs decrease the bias on the estimation of new observations. Furthermore, the averaging of many realizations of trees decreases the variance on the outcome (Alpaydin, 2014).

This method does not predict out of sample observations. Consequently, it requires an input that best describes the behavior of the data. Moreover, many algorithms build the forest by raising trees with different variables.

2 MATERIALS AND METHODS

The data under analysis is composed of an artificially generated set and a collection of economic indicators. The artificial set contains nine variables that are dependent on time and depict various scenarios. This set contains 100 time series intervals of 90 periods and the variables describe the following trends: linear; multiplicative; white noise; autoregressive process; moving average process; changing growth; weekly frequency; monthly frequency; and, varying frequency. In contrast, the real world data consists of indicators containing unknown behavior.

The analysis of the artificial set allows the implementation of a methodology to evaluate the model parameters and data stationarity. As a result, this framework extracts the most information from the unknown behavior of the financial indicators. Furthermore, the analysis divides both sets in 90 days intervals of which the last 10 days correspond to the forecasting window. This is an ideal time series, however, differencing operations will decrease this interval. Hence, the forecasting window is variable and follows the proportion 10/90, depending on the size of the interval. Lastly, this methodology follows four stages: i) data pre-processing; ii) forecasting strategy; iii) model selection and parameter estimation; as well as, iv) error calculation.

The scripts implemented in this approach are publicly available on the following GitHub repository: https://github.com/ronnierp/T61.5910TSA.

2.1 Data pre-processing

The stationarity of the data is important for the ARIMA model. In this sense, the analysis of the indicators is restricted only to those that fulfill the stationarity requirement. Consequently, the pre-processing scheme performs the Augmented Dickey-Fuller (ADF) test at different stages. The first evaluation happens with the raw data. If the variable fails the test, then the ADF becomes a stopping criterion for the differencing operation. In other words, this iteration occurs up to five times and the significance threshold for the test statistic is equal to 1%. Notwithstanding, economic indicators usually become stationary after no more than two iterations.

The differencing operation presents the problem of under-differencing or over-differencing. Therefore, the differenced variable should not have variance greater than that of the previous iteration. Hence, an iteration that rendered a variable stationary, but increased its variance is reverted. This increase in variance is a sign of over differencing, however, an extra MA term might correct it. At the same time, the introduction of an AR term mitigates under-differencing (Nau, 2015).

Another important aspect of stationary data is the absence of trends. In this regard, if the differencing operation was unsuccessful to stabilize the variable, then it undergoes a trend test. This procedure removes any constant term or linear trend before resuming the stationarity routine described above.

However, the pre-processing stage still requires the analysis of the autocorrelation and partial autocorrelation functions. This analysis establishes if the lags of the variable are close to zero and if they have a decreasing trend, since stationary data should have high dependence only on the initial lags. In this sense, the test comprises of evaluating the number of lags outside the confidence interval of the variable as well as the magnitude of its decrease. If more than 40% of the lags fall outside of the interval, then the variable is classified as non-stationary. This threshold becomes 60%, if the variable showed signs of over-differencing, in order to compensate the premature finish of the differencing operation.

The last procedure in this stage is to standardize the data, thus subtract the mean and divide by the variance.

2.2 Forecasting strategies

This paper pursues three forecasting strategies, that is, 1-step, 2-steps, and h-steps (Tsay, 2010). The first corresponds to forecast one point and feed it back to the training data for the next forecast prediction. This strategy might lead to an increased error at the end of the forecasting window, since forecasted steps might rely on inaccurate predictions (Bontempi, 2013). Figure 2.1 presents this approach, emphasis is given to the 10 day period of the forecasting window which depicts the uncertainty of the first forecast in light gray.

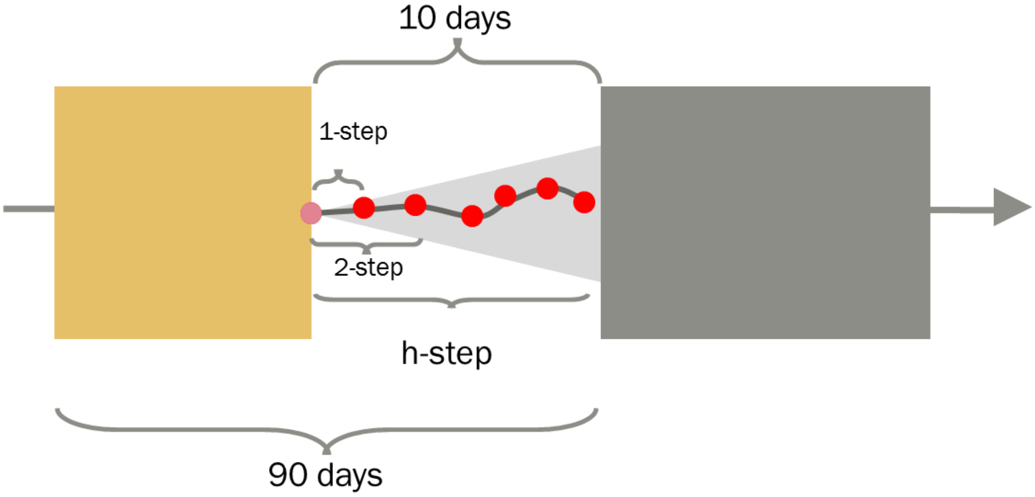

Figure 2.1 – Forecasting Window and Strategies.

The next approach relies on a similar methodology, however, it tries to circumvent the compounding error by retroactively adding every other forecasted prediction to the training set. This corresponds, to forecast two steps ahead of the current time at each prediction. Finally, the h-steps approach does not depend on changes to the training set. The operation predicts the entire forecasting window at once.

2.3 Forecasting models and parameter estimation

This analysis compares univariate and multivariate models. Furthermore, the univariate category is composed of three types of models: AR, MA, and ARIMA. In contrast, the multivariate class comprises VAR and Random Forest.

2.3.1 Autoregressive Integrated Moving Average - ARIMA

The univariate models have three parameters (p, d, and q) and they correspond to AR(p), I(d) and MA(q). However, the stationarity requirement already provides “d”, consequently the estimation depends only on evaluating the pure model coefficients.

The answer to this problem lies in evaluating the lag dependency of the Partial Autocorrelation Function (PACF) and Autocorrelation Function (ACF). The first returns the AR coefficient and the second the MA. Therefore, the procedure consists of determining the argument that minimizes those functions within the first five lags, given that it lies inside the confidence interval. As a result, this procedure sets the maximum threshold to five for each coefficient.

The combination of pure model parameters gives the full ARIMA model. Nonetheless, the parsimony principle advocates for the model selection with the least amount of parameters (Coghlan, 2015). However, the performance of the models were also taken into consideration. In this regard, the data is divided in two sets: training and validation.

The training set determines the pure model coefficients, in this sense the limits to the possible combinations of the mixed approach (ARMA). Consequently, the validation set evaluates all combinations. Nevertheless, pure models have precedence over mixed models. Furthermore, the Akaike Information Criterion (AIC) is the estimate of the loss of information on the generalization error. Thus, the coefficients are those that minimize the AIC score.

An extra consideration on the parameter estimation lies on the choice of training and validation set. This is noteworthy, because of the particular characteristic of the ARIMA models, which is the time series lag dependency. In this sense, the division of the sets occur at a defined point and not at random. Therefore, the training data comprises the data points that span 75% of the initial interval while the remaining data points that lie immediately before the forecasting window constitute the validation set. In order to further clarify this point, it is important to note that the data is stationary and this implies constant mean and no trend, consequently the entire set should retain these traits.

There is parameter estimation with every move of the forecasting window, since the prediction of the previous iteration expands the training set. Moreover, all strategies also require parameter estimation. Finally, the standard forecasting methodology for ARIMA models is the minimization of the sum of squared error. This report implemented the forecasting functions given by the StatsModels Python package, especially the section related to Time Series Analysis.

2.3.2 Vector Autoregressive – VAR

This model follows similar rationale to the methodology described for ARIMA, but it carries a single parameter. This coefficient corresponds to the number of lags across variables that contributes to the result, in addition its upper limit is also set to five. Furthermore, the order selection relies on the evaluation of the Akaike Information Criterion over the validation set as well as new parameter estimation at each move inside the forecasting window.

2.3.3 Random Forest

The random forest regression problem averages the result 100 decision trees, where each tree contains enough decision rules to produce pure leaves. Furthermore, due to the absence of out of sample predictions a weighted average serves as input. The average depends on the last eight data points and assigns an increasing weight to points closer to the forecasting window. In this regard, the point immediately before the forecasting window has a weight equal to eight while the last amounts to one.

The selection of eight data points was motivated by the analysis of the generalization error over a range of weights. This value corresponded to the approximate location of the lever axis on the elbow plot of the generalization error versus weight size.

The final information obtained from the random forest is the importance of each feature to the forecast. The retrieval of this value followed the default output of feature contribution on the random forest regression algorithm provided by Scikit-learn.

2.3.4 Mean

The forecast given by the variable mean is an uninformative prediction, since it does not say whether the indicator increases or decreases. However, it provides a benchmark to compare the models. In this sense, all predictions by this model consist of the average value of the indicator in the training set.

2.4 Generalization Error Calculation

The standard method implemented in this paper is that of the Sum of Squared Errors. This corresponds to the sum of the squared differences between forecast and the true underlying value of the time series. Furthermore, this study reports the median error as well as stacked histograms of the errors corresponding to the first, fifth and final move inside the forecasting window.

The summarization of the error of each move by the median presents an estimate untainted by outliers. Thus, every model and strategy report the median of each step allowing direct comparison against the benchmark as well as across model/strategies. Furthermore, the bars on the histograms carry great comparative power, since the individual bars represent the proportion of contribution given by each model to the given bin.

The final remark about the error estimate corresponds to the comparisons against the benchmark. This estimate is the baseline for predictions, in this sense the error of a model divided by the benchmark gives a better understanding of its predictive power than only the sum of squared errors. For instance, a prediction closer to the true underlying result leads to a lower error than the benchmark. Consequently, the ratio between this generalization error and the benchmark should yield a number lower than one. In contrast, worst predictions could greatly depart from one. As a result, this study presents the error as the ratio between the model performance and the benchmark.

3 RESULTS

This section presents data on the generalization error dependent on strategy and model. Furthermore, the results are divided according to the type of data: simulated and financial indicators.

The latter also contains an analysis of feature importance while the former presents the choice of number of weights for the input of the random forest algorithm.

3.1 Artificial Data

Figure 3.1 presents the error associated with the choice of weighted average terms.

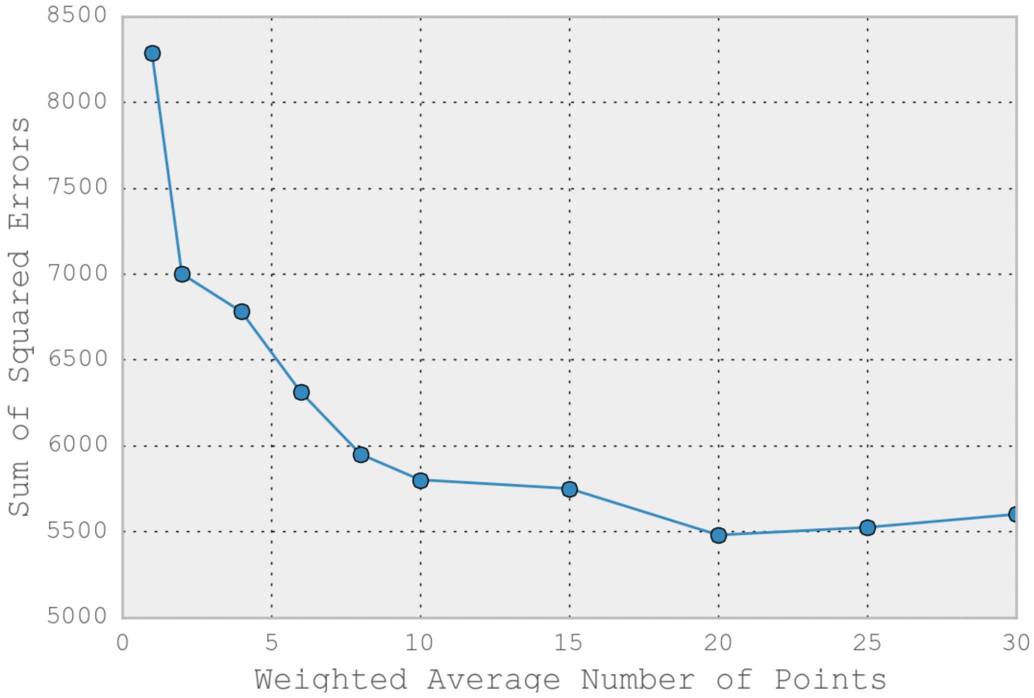

Figure 3.1 – Sum of Squared Errors for varying weighted average terms.

The following table shows the influence of forest size in the model accuracy.

Table 3.1 –Sum of Squared Errors for the Random Forest model with varying forest size.

Strategy	Forest Size	
	10 trees	**100 trees**
1-step	5433	5402
2-step	310	301
h-step	732	642
Total	**6476**	**6346**

3.1.1 Overall Prediction Scores

Table 3.2 summarizes the error inherent in the forecasts of the artificial data set according to step, strategy and model.

Table 3.2 – Generalization Error According to Model and Strategy (Simulated Data - SD).

Step	Mean	1-step			2-steps		h-steps		
		ARIMA	VAR	RF	VAR	RF	ARIMA	VAR	RF
1	51.65	0.524	0.395	0.833	0.915	1.017	0.524	0.395	1.568
2	46.77	0.735	2.319	2.451			0.735	1.011	2.251
3	45.33	0.505	4.044	4.582	1.622	1.467	0.514	0.720	0.847
4	68.37	0.477	4.506	4.473			0.593	0.771	1.002
5	56.95	0.775	9.120	7.579	2.236	1.210	0.782	1.186	1.296
6	40.05	0.568	15.712	14.525			0.627	1.318	1.732
7	64.69	0.495	12.680	11.696	1.820	0.820	0.633	1.019	1.437
8	43.86	0.675	20.772	20.123			0.751	1.088	1.176
9	54.34	0.521	18.253	18.426	2.082	1.372	0.587	1.091	1.327
10	62.93	0.502	17.163	17.612			0.622	0.919	1.258

Sum of Squared Errors for the mean (benchmark) and error ratios, with respect to the benchmark, for all others.

3.1.2 1st step

Figure 3.2 depicts the 1-step strategy SSE histogram, while Figure 3.3 and Figure 3.4 represent the error in the 2-step and h-step approaches, respectively.

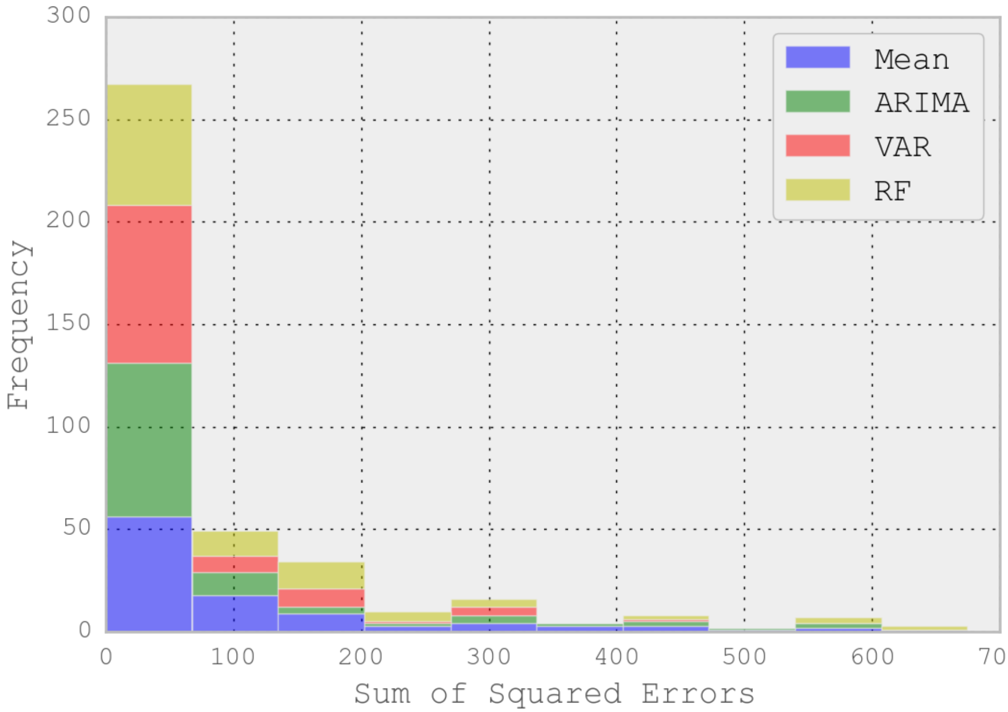

Figure 3.2 – 1-step histogram of SSE (SD, 1st step).

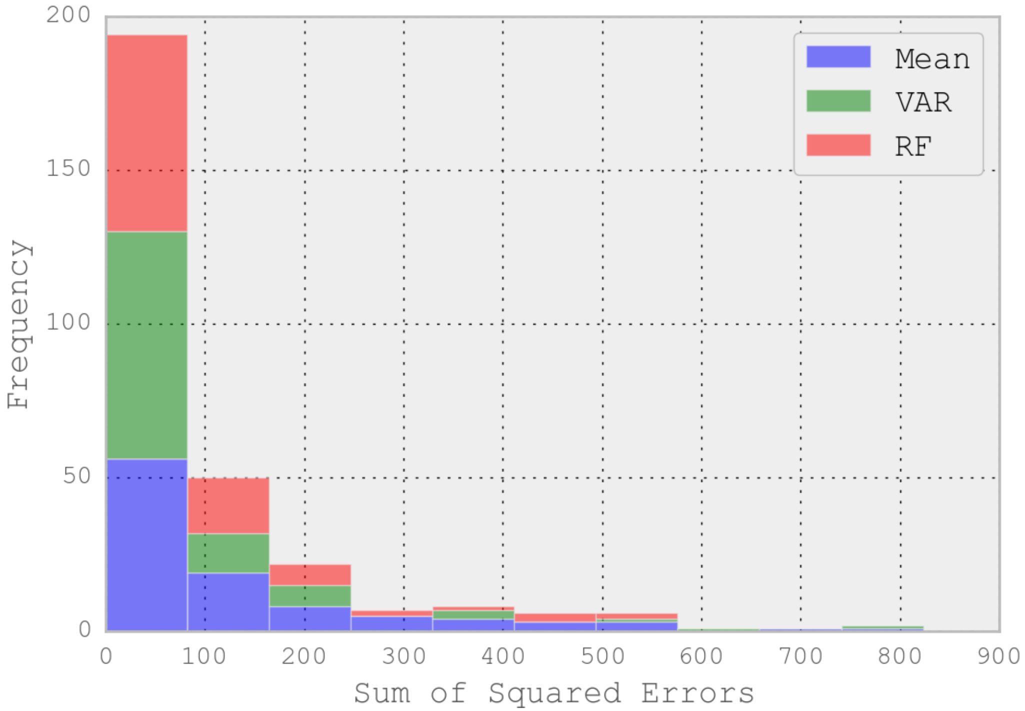

Figure 3.3 – 2-step histogram of SSE (SD, 1st step).

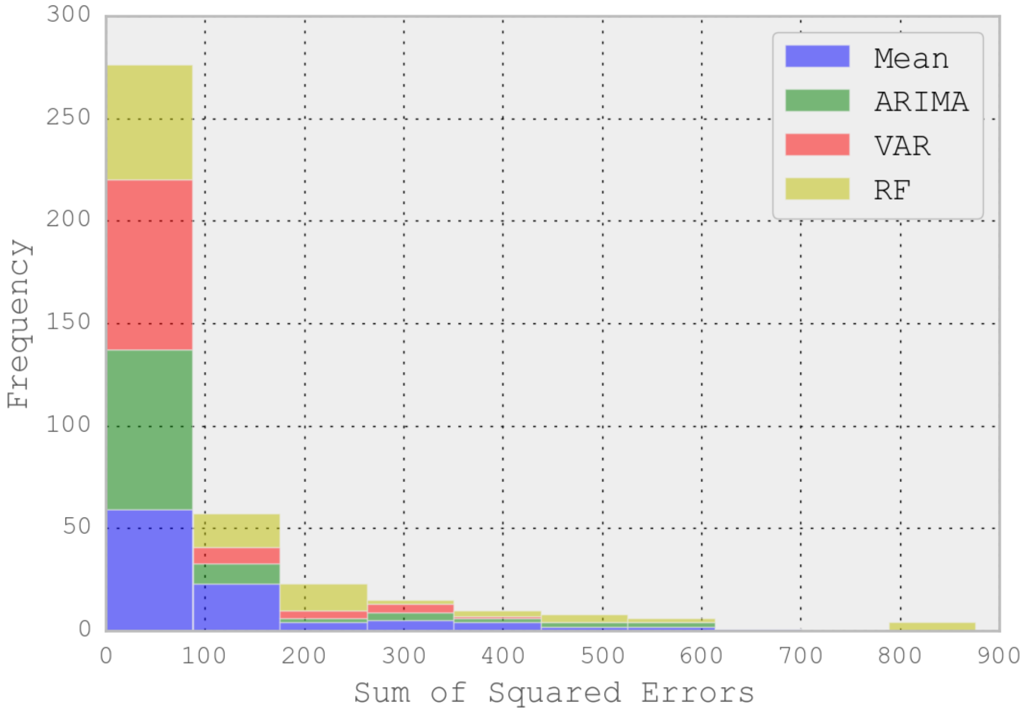

Figure 3.4 – h-step histogram of SSE (SD, 1st step).

3.1.3 5th step

The plots containing the histograms relative to the fifth step concern only the 1-step (Figure 3.5) and h-step (Figure 3.6) strategies.

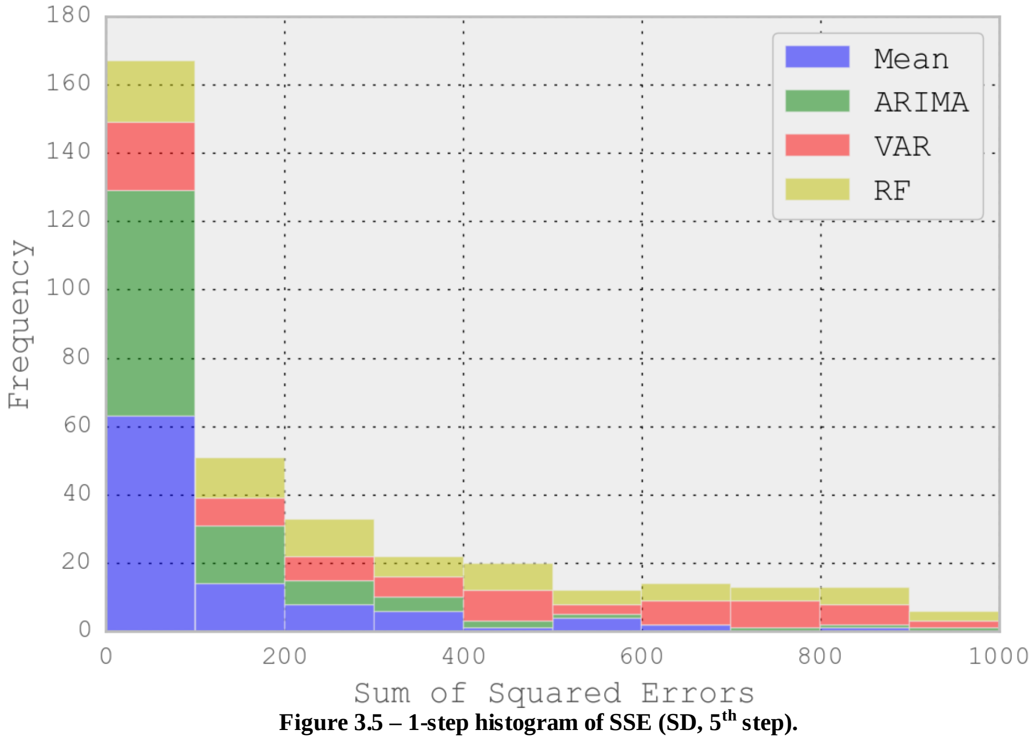

Figure 3.5 – 1-step histogram of SSE (SD, 5th step).

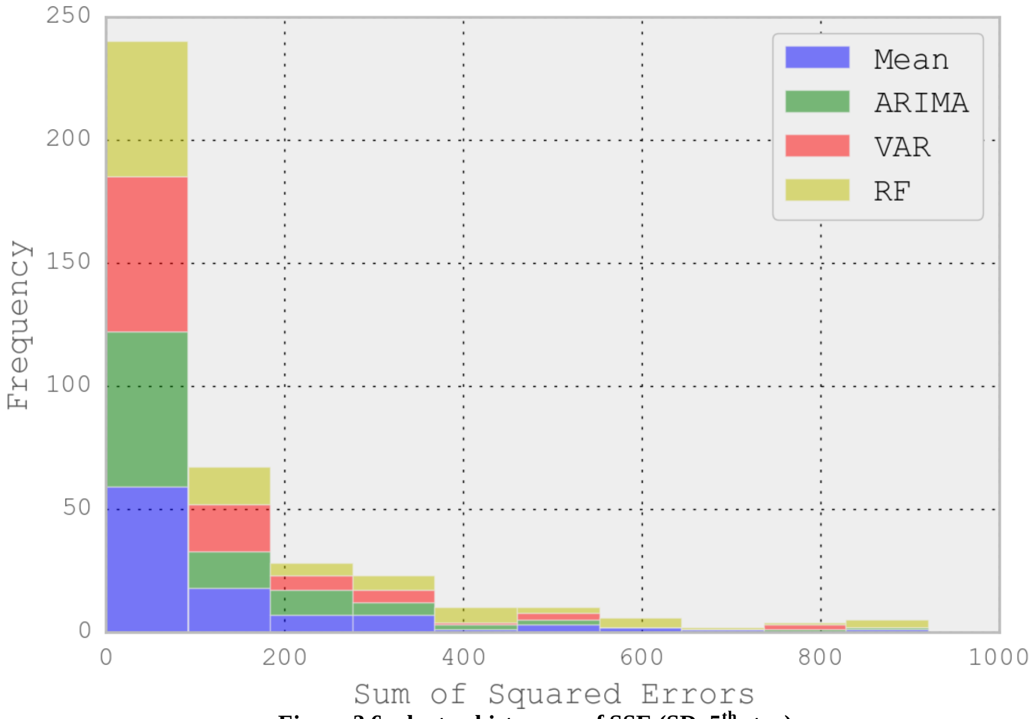

Figure 3.6 – h-step histogram of SSE (SD, 5th step).

3.1.4 10th step

The tenth step comparison presents histograms for all strategies. The plots on Figure 3.7, Figure 3.8, and Figure 3.9 represent the following strategies: 1-step, 2-step, and h-step.

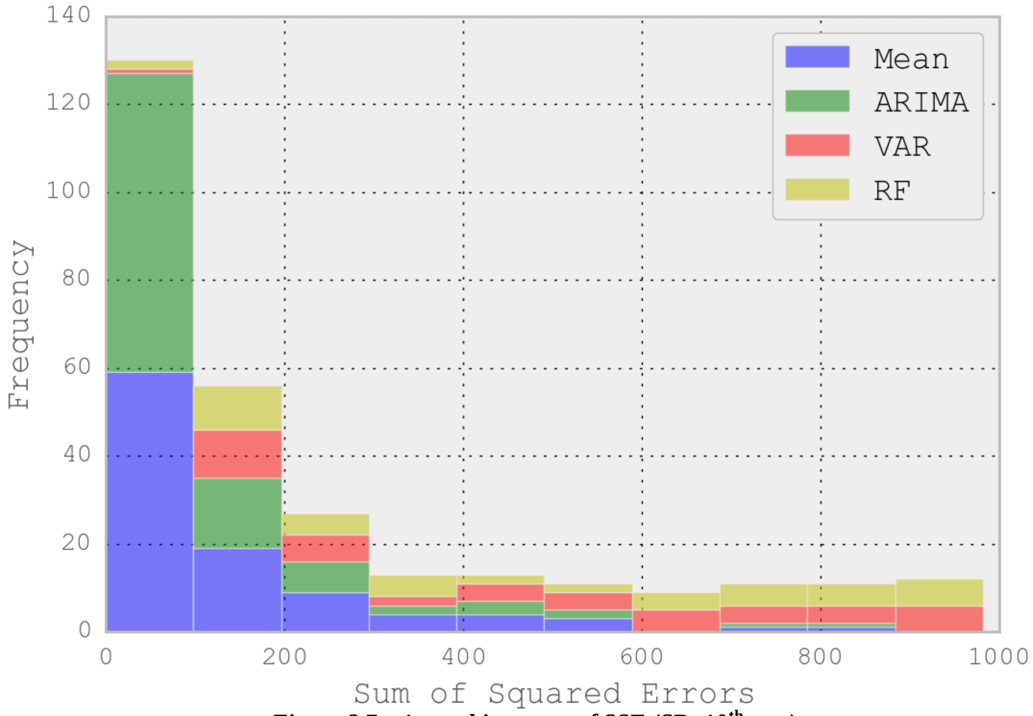

Figure 3.7 – 1-step histogram of SSE (SD, 10th step).

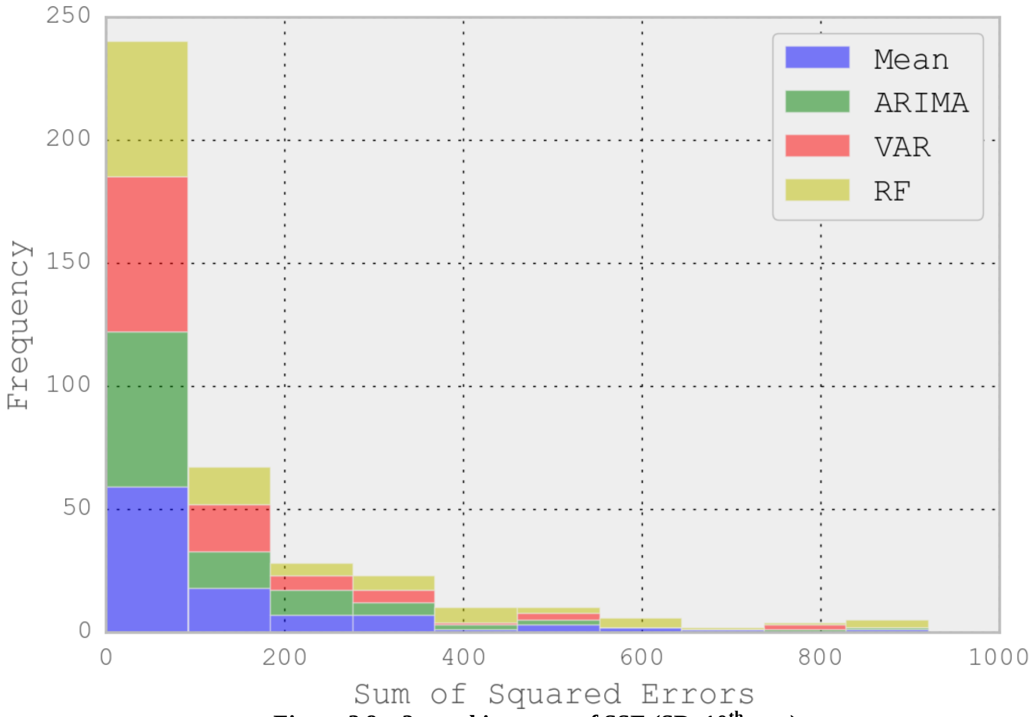

Figure 3.8 – 2-step histogram of SSE (SD, 10th step).

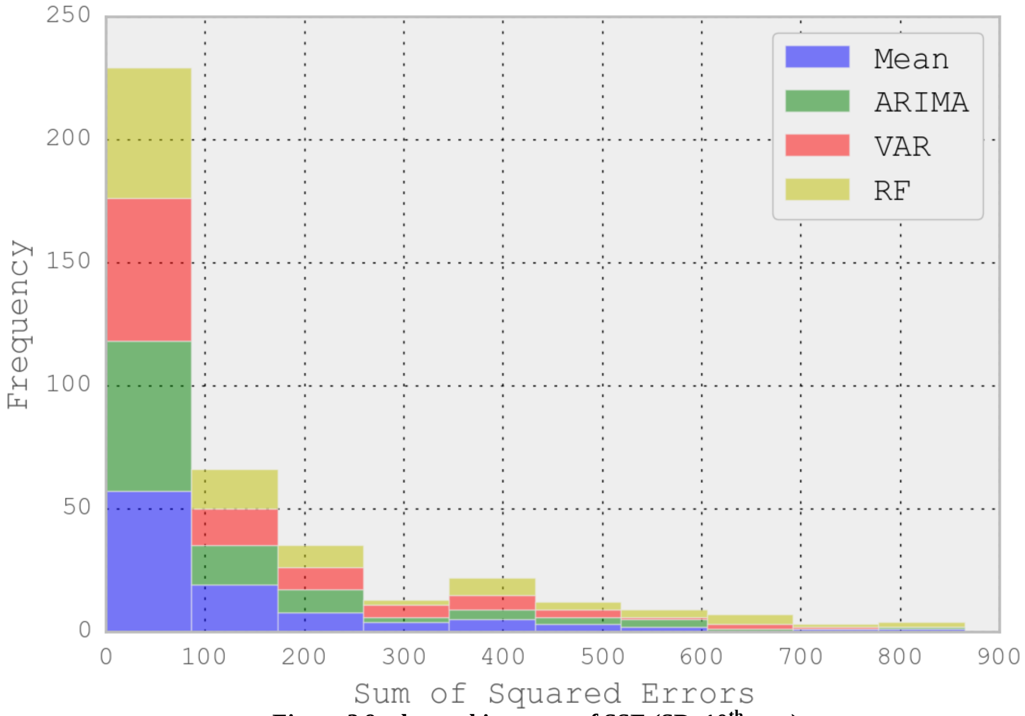

Figure 3.9 – h-step histogram of SSE (SD, 10th step).

3.2 Financial Indicators

This section follows the same organizational pattern of the simulated data results.

3.2.1 Overall Prediction Scores

The table below contains the information of all models and strategies divided according to the move/step inside the forecasting window. The benchmark forecast is the sum of squared errors while the other models present the ratio between their errors and those from the benchmark.

Table 3.3 – Generalization Error According to Model and Strategy (Financial Indicators - FI)

Step	Mean	1-step			2-steps		h-steps		
		ARIMA	VAR	RF	VAR	RF	ARIMA	VAR	RF
1	185003	0.899	11.133	0.897	11.133	0.954	0.899	11.133	1.418
2	189588	0.939	36.151	2.174			0.927	17.398	1.431
3	193014	0.926	50.518	3.466	17.305	0.983	0.960	13.967	1.235
4	188763	0.948	86.649	4.479			0.951	14.263	0.987
5	147103	0.930	170.847	7.204	259.050	0.906	0.961	16.926	1.189
6	144425	0.849	209.664	9.024			0.931	28.515	1.286
7	143451	0.955	204.403	9.973	309.875	0.867	0.965	28.115	1.206
8	126892	0.980	402.665	12.954			0.985	42.489	1.354
9	174767	0.936	324.282	11.347	5.74E+03	1.426	0.984	33.454	1.226
10	184596	0.909	308.619	11.574	9.36E+05	1.477	0.913	28.559	1.155

3.2.2 1st step

Figure 3.10 - Figure 3.12 illustrate the behavior of the error for the first step.

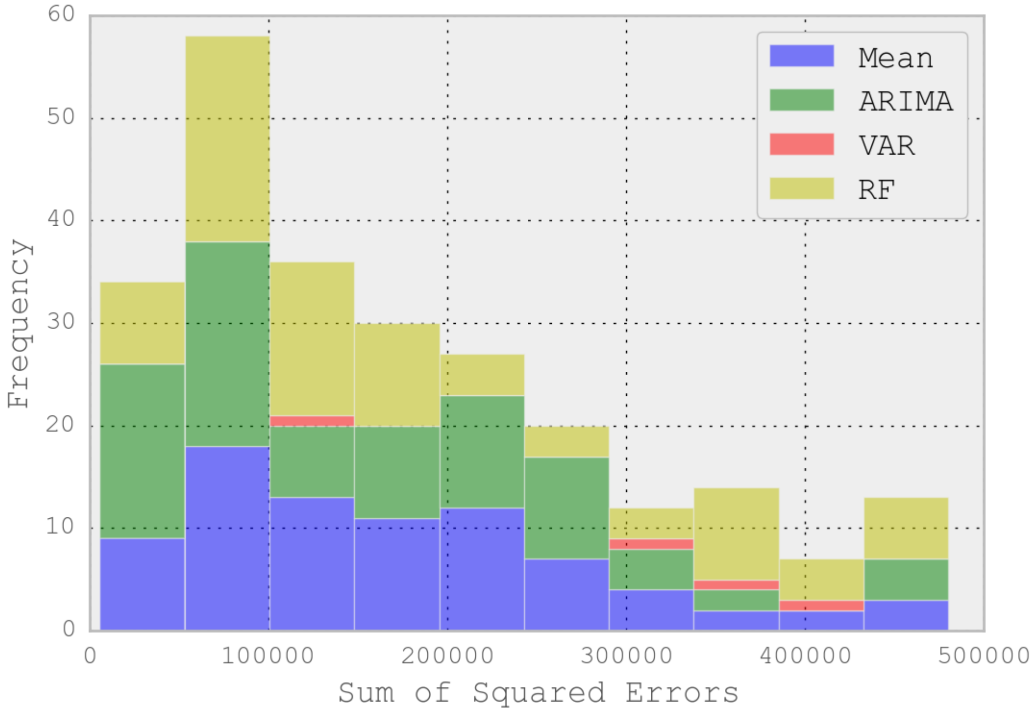

Figure 3.10 – 1-step histogram of SSE (FI, 1st step).

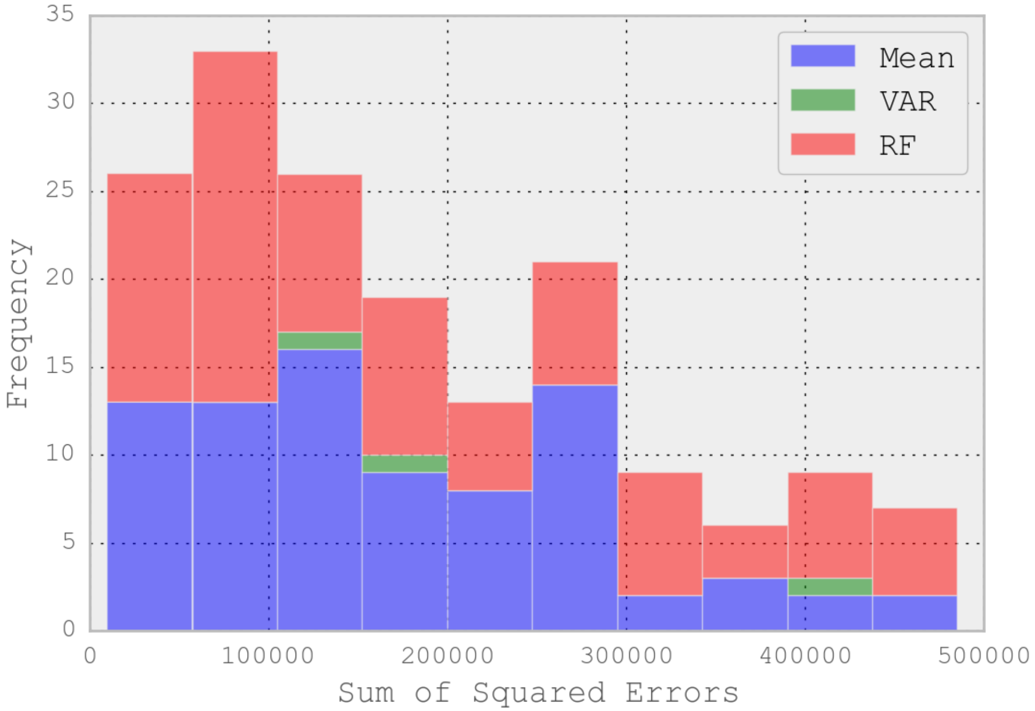

Figure 3.11 – 2-step histogram of SSE (FI, 1st step).

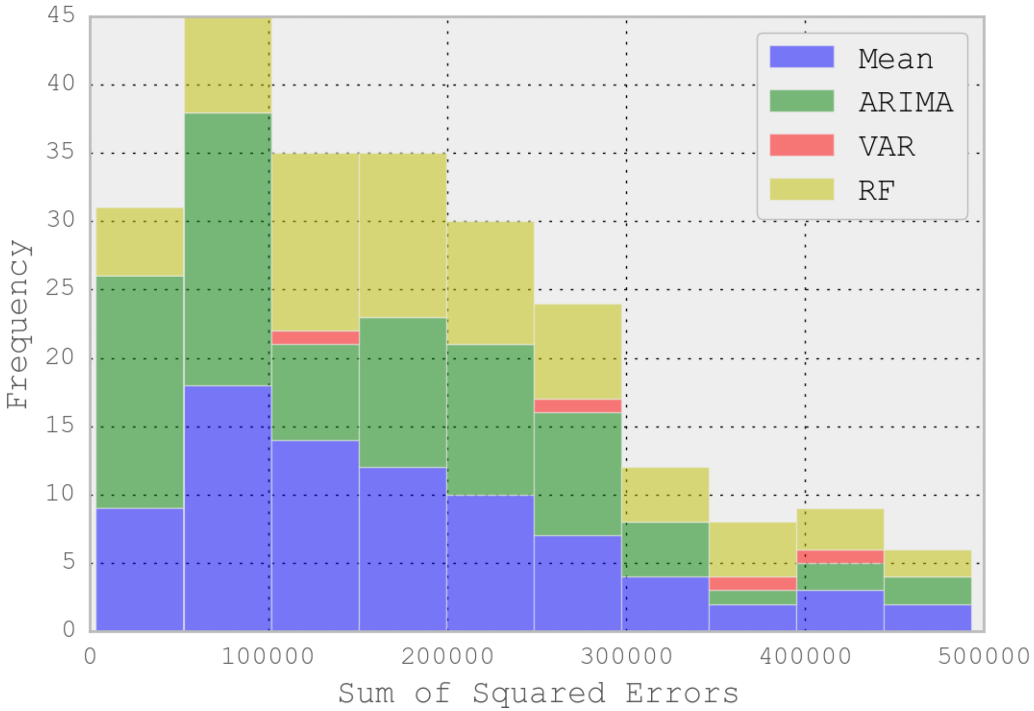

Figure 3.12 – h-step histogram of SSE (FI, 1st step).

3.2.3 5th step

Figure 3.13 and Figure 3.14 display the 1-step and h-step strategy results.

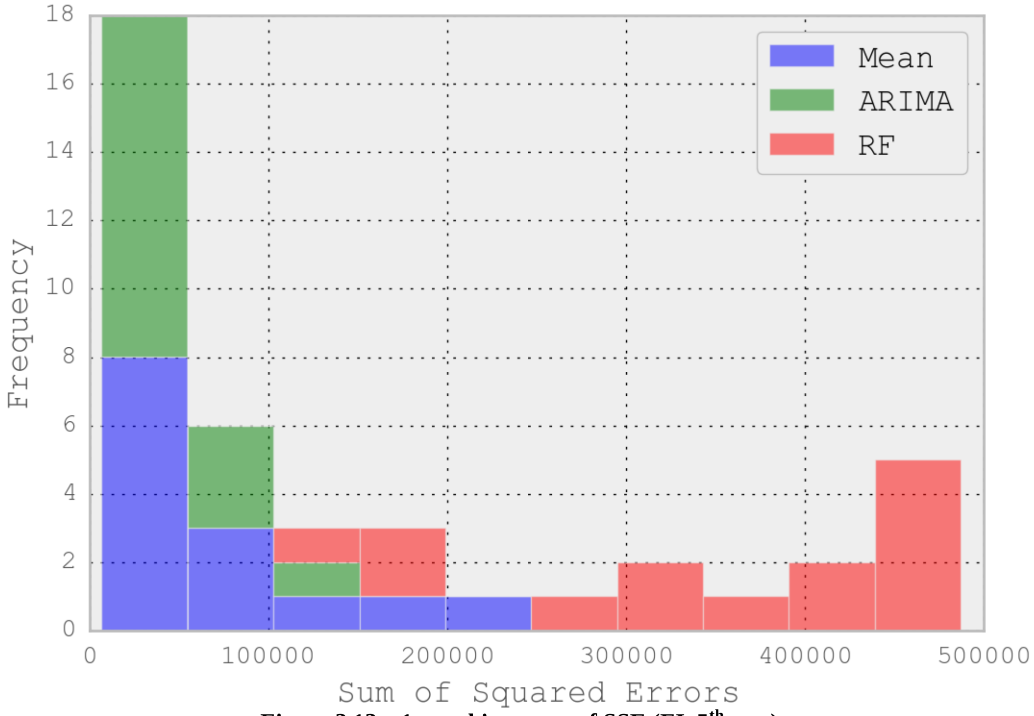

Figure 3.13 – 1-step histogram of SSE (FI, 5th step).

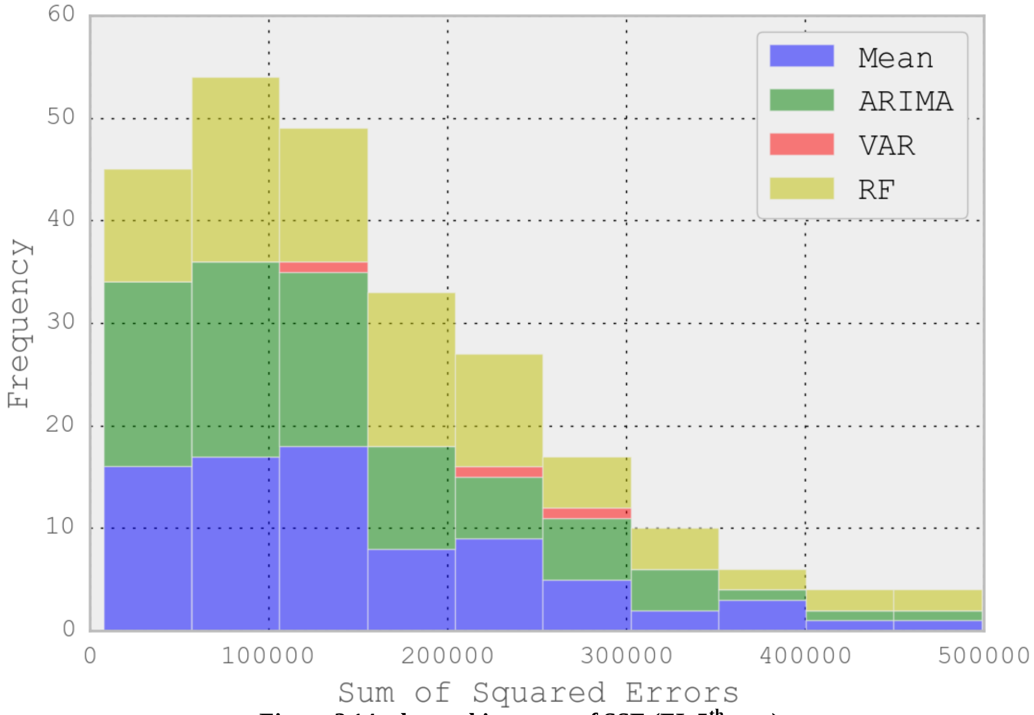

Figure 3.14 – h-step histogram of SSE (FI, 5th step).

3.2.4 10th step

The 2-step strategy yielded no forecasts, hence Figure 3.15 and Figure 3.16 represent the 1-step and the h-step approaches.

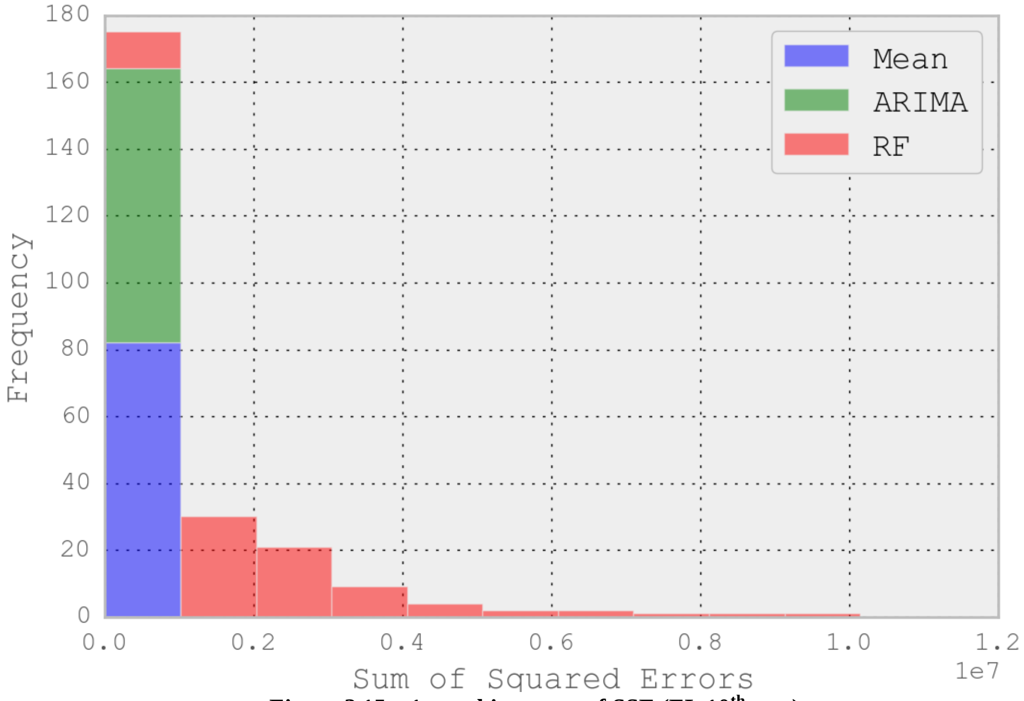

Figure 3.15 – 1-step histogram of SSE (FI, 10th step).

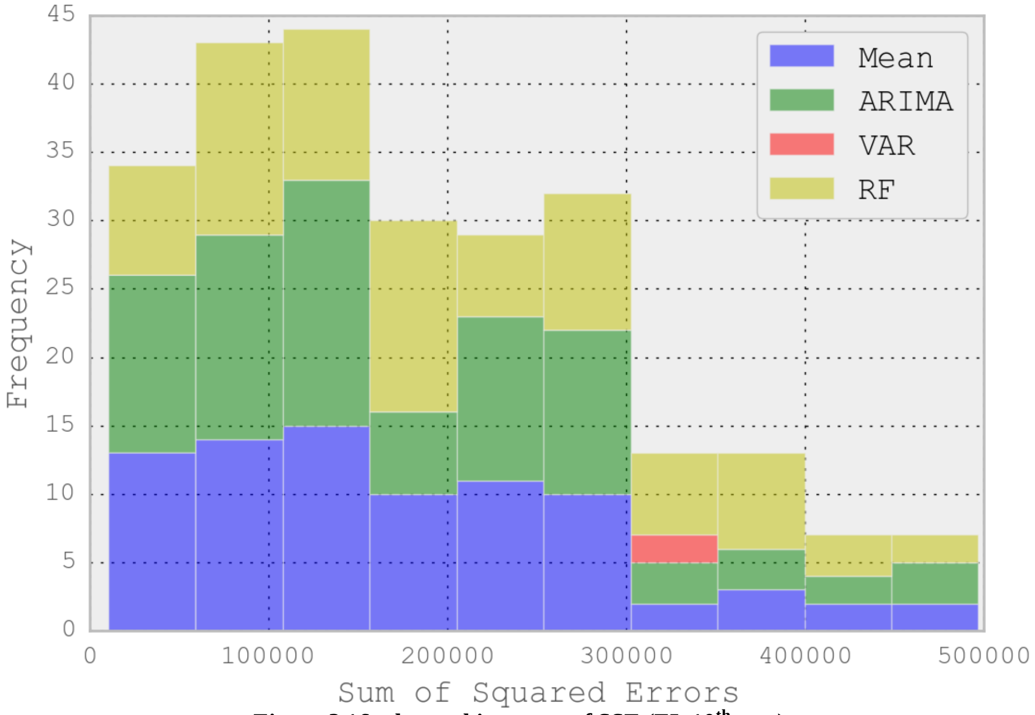

Figure 3.16 – h-step histogram of SSE (FI, 10th step).

3.2.5 Feature Importance

The table below lists the rank and the name of the features that contributed the most to the forecasts of the random forest model.

Table 3.4 – Feature Importance

Rank	Financial Indicator
1	stock_index_nyse
2	stock_index_russell_3000
3	stock_index_spx_500
4	stock_index_sp400_midcap
5	stock_index_sp100_options
6	industry_index_nasdaq_industrial
7	stock_index_volatility
8	stock_index_nasdaq
9	stock_index_nyse_arca
10	currency_to_usd_chf
11	industry_index_nasdaq_bank
12	industry_index_nasdaq_transportation
13	currency_to_usd_aud
14	commodity_vanguard_precious_metals_and_mining_inv
15	commodity_silver

4 DISCUSSION

The main goal of the project was to evaluate different forecasting models, namely two classes: univariate and multivariate. In addition, this report evaluated multiple prediction strategies and the influence of the size of the forecasting window on the accuracy of the forecast.

Furthermore, all models depended on parameter estimation. Thus, the iterations inside the forecasting window estimated parameters for the univariate and the VAR models. In contrast, the random forest parameters were selected based on tests on the artificial data.

4.1 Artificial Data

An analysis of the generalization error in the random forest predictions, with respect to the size of the weight vector, determined the number of terms in the weighted average input. Figure 3.1 depicts this behavior. Furthermore, this plot shows a typical “elbow” shape that corresponds to a sharp decrease in the error for a small number of weighted average terms, followed by small decrements after a threshold. The current scenario showed this behavior for vector sizes 2 – 20 and an approximate location of the axis in points 8 – 10. Hence, the selected vector size contains eight terms. This vector gives variable contribution to the inputs, varying from one to eight. In other words, the weights give higher contribution to values immediately before the forecasting period and decrease up to the eighth furthest data point.

Moreover, the depth of the decision trees and their number directly influence the predictive bias and variance. Therefore, the trees are extended until pure leaves, in order to reduce the bias to the highest extent possible. Moreover, in an effort to minimize the variance the forest size was set to 100. Table 3.1 shows the SSE of forests with 10 and 100 trees. It is possible to see that the increase in size leads to a decrease in error. It is worth mentioning that an increase in the number of trees comes with computational costs, hence larger forests were not pursued to prevent large time complexities, given the iterative behavior of the strategies and overall computational cost of the implementation in this study.

Table 3.2 shows the overall behavior of the forecast error when compared to the benchmark. This standard gives an uninformative forecast, since it does not tell whether the indicator rises or falls. Hence, favorable predictions are those, which return a value lower than or equal to one. The latter bound equality to one includes the possibility that the indicator, for the given time point, lies in the mean of the time series. However, for the purpose of this paper, the convention when addressing the relative error assumes low error when the ratio is closer to zero than to one.

The first step into the forecasting window shows very good results for the predictions of all models. This value reaches its lowest level in the VAR model followed by ARIMA and random forest. Overall, ARIMA outperforms all models in the 1- and h-step strategies while VAR produces more accurate results than a random forest in the 2-step strategy. The ARIMA model is not included in the 2-step strategy due to the choice of algorithm that poses a strict dependence on the first time point before the forecasting period.

The proportion of accurate predictions is better explained by the stacked histograms. Figure 3.2 - Figure 3.4 show high frequency of accurate predictions near zero and a long tail towards higher SSE. The contribution of the random forest model in the 1- and h-step strategy (Figure 3.2 and Figure 3.4) is the smallest around the mean, while on the 2-step approach, it is comparable to that of VAR. Overall, the plots show similar to larger bars when compared to the prediction by the mean, which corresponds to a high number of successful predictions.

The fifth step does not include the 2-step strategy, because not all time series were evaluated at this step by this strategy. The prediction depended on the size of the forecasting window, hence odd numbered windows are not included. The performance of the ARIMA model is superior to all others in the 1-step approach (Figure 3.5), but equivalent to all other models in the h-step (Figure 3.6). However, the error on the test set remains the lowest among all models (Table 3.2).

At this time point, the 1-step approach starts to exhibit strong influence on past inaccurate predictions (Table 3.2). In other words, this strategy shows an increasing median error on the VAR and random forest models that is the result of compounding error. This behavior is not present in the h-step strategy, which displays small variance.

The compounding effect reaches its maximum at the tenth step, as depicted in Table 3.2. In this scenario, only the ARIMA model gives a successful prediction. This behavior is also present in the 2-step approach, but to a much lesser extent. However, the predictions on this approach are higher than those of the benchmark.

The distribution of the sum of squared errors has the same characteristic behavior of the previous steps, however, there are changes to the contribution of each model. For example, Figure 3.7 shows the burden of the compounding effect on the multivariate models, hence one notices that the ARIMA model gives the largest contribution to accurate predictions. In contrast, the 2- and the h-step approaches (Figure 3.8 and Figure 3.9) have similar histograms with more evenly distributed contributions among models.

The analysis of the artificial data set showed that the compounding effect plays an important role in the 1-step strategy. Furthermore, the 2-step approach considerably decreases this effect, which ultimately translates in predictions that are more accurate. In sum, the univariate models outperform the multivariate models in the iterative strategy and VAR has a greater proportion of accurate predictions than the random forest in the 2-step approach. Finally, ARIMA also gives better predictions in the direct h-step strategy.

The VAR model had similar performance to the random forest and at some points was slightly superior. However, a single AR process might not be enough to explain the data, even though there are lagged terms across variables. The lack of an MA parameter might be an indicative of its inferior performance when compared to the ARIMA model. Lastly, within the multivariate models VAR shows good promise.

Finally, the random forest forecast only showed good performance on the first prediction of the 1-step approach. This is an indication that just the weighted average might not be enough to capture the underlying behavior of the data. In this sense, the analysis of the financial indicators also covers feature importance.

4.2 Financial Indicators

The behavior of the models in this dataset shows peculiar characteristics, especially for the Vector Autoregressive methodology. This model consistently failed to provide reliable predictions and its contribution to the plots in Figure 3.10 - Figure 3.16 are negligible. This behavior was unexpected, given the performance of this approach on the artificial dataset. However, the indicators consist of data that are more intricate and it is very likely that only AR terms are not able to fully describe the indexes. In this sense, similar to the case of the artificial data set, this model would benefit from the addition of Moving Average terms. This claim finds basis on the overall performance of the ARIMA model in this set. The latter approach performed in the lower range of errors in most scenarios.

An alternative to the addition of MA terms would be to lag variables that contribute the most to the data variability. An example of this approach consists of the variables presented on Table 3.4. This list of variables is an initial attempt at feature selection, therefore, the model might behave differently if configured to forecast based on this filter. Nonetheless, the VAR model proved unsuitable for analysis of financial indicators, thus this study presents detailed discussion only on the remaining models.

The benchmark of the financial indicators has a sum of squared errors much larger than the baseline of the artificial data set. The reason for this discrepancy lies in the number of predicted variables, since the sum of squared errors is an additive indicator. However, the ratio between the benchmark and the forecast from different models still carries the same interpretation. Consequently, a ratio lower than or equal to one consists of a forecast that shows an improvement over the prediction of the benchmark while values higher than one correspond to lower accuracy.

The first step in the forecasting window had similar results across models. The 1-step strategy yielded equivalent predictions in the ARIMA and RF approaches. In contrast, the h-step methodology showed favoritism for the ARIMA model while the 2-steps strategy for the random forest

Figure 3.10 depicts the performance of all models in the 1-step strategy. This plot shows a higher concentration of points towards the lower region of the generalization error, which denotes consistent performance across models. On the other hand, Figure 3.11 shows the distribution of errors in the predictions of the random forest compared to the benchmark. It is noticeable that this strategy yielded fewer predictions in the range of 0 – 5.10^4, moreover, the distribution from the random forest showed slightly higher counts in the lower region than those from the benchmark. However, these are marginal gains, consequently the ratio of the median error (0.954, Table 3.3) is a true description of the model, since in the set of forecasted variables most where uninformative. Lastly, the h-step strategy, depicted in Figure 3.12, clearly shows the ARIMA model outperforming the random forest. This behavior is also evident in Table 3.3, since the median ratio of the first is equal to 0.899 while the latter amounts to 1.418.

The ARIMA model showed the most accurate predictions in the fifth step of the forecasting window in the 1- and h-steps approaches. Nonetheless, the best forecast came from the random forest in the 2-steps strategy. Table 3.3 summarizes this scenario.

In addition, Figure 3.13 illustrates the compounding effect of the error on the predictions. For instance, the overall number of predictions in the range of $0 – 5.10^4$ is very low. Moreover, the random forest model is incapable of producing accurate predictions. As a result, its distribution appears skewed to the left. On the contrary, when a prediction was possible, the ARIMA model had a satisfactory performance yielding a distribution that resembles an exponential function.

The h-step strategy unlike the previous approach does not rely on forecasting based on previous predictions and as such yielded better performance. For example, Figure 3.14 presents distributions skewed to the right for all models. In addition, the number of predictions is higher than that on Figure 3.13. Moreover, the ARIMA model yet again showed equivalent to superior performance in this approach.

The deviation from the true value of the indicators reaches its highest amplitude in the final move of the forecasting window, namely on the 1-step approach. In order, to illustrate this dependence the axis on Figure 3.15 was extended to include all predictions provided by the models. This plot shows the full extent of the compounding effect, especially on the random forest model. As a result, apart from the first bin, this plot depicts only false predictions. This behavior is quite the opposite of the h-step approach (Figure 3.16). However, the performance of this model is still inferior to ARIMA in both scenarios. Hence, Table 3.4 shows a list of features with the highest contribution to the forecasting decisions.

The evaluation of the random forest model yielded the percentage contribution of each feature to the prediction. As a result, all variables were tracked at every evaluation of the random forest model. In this sense, Table 3.4 presents the intersection between the information obtained from all strategies (1-, 2- and h-steps). Furthermore, this table shows a high proportion of indexes that summarize the general state of the market (collective) rather than individual indicators. For instance, the Russell index depicts the change in a collection of prices from stocks of US companies while the S&P series (SP on the list) presents the volatility of the market. Moreover, it is also worth mentioning that this list contains only a small number of commodity indicators, that is, the price of silver and the more general indicator for precious metals. In addition, the US dollar plays an important role in the selection of currency indicators with especial emphasis on the conversion to Australian Dollars and Swiss Francs.

Overall, this study included 258 indicators of which a varying number of 30 – 50 were stationary. Furthermore, an analysis of feature importance shortlisted 15 indicators that gave major contributions across strategies. In this sense, further studies on this topic would greatly benefit from modeling the forecasts based on these features. This includes the individual features as well as combinations of the existing variables in new features.

In summary, the performance of the methods of this study in the given data set was successful for the random forest and the ARIMA models. In contrast, the VAR approach proved not enough to describe the behavior of the data. Furthermore, the use of three strategies to step through the forecasting window showed the influence of compounding error on the predictions. In this sense, the future direction of this work should focus on improving the forecast of the h-steps strategy. This implies that the most trivial case of predicting the next period of the time series would also be covered by this strategy. Moreover, this direction greatly decreases the computational cost of the study, since the selection of parameters occurs only once for each forecasting window. Lastly, the univariate model outperformed the random forest, however, the latter could benefit from further investigation based on the shortlisted features.

5 CONCLUSION

This study evaluated the forecast accuracy of univariate and multivariate models applied to time series data, namely, financial indicators. The models relying on a single variable consisted of the Autoregressive Integrated Moving Average approach and its pure forms (AR and MA) while the models dependent upon multiple variables comprised of Vector Autoregressive and Random Forest. Furthermore, the study evaluated three strategies to step through the forecasting window. This methodology proved successful in forecasting variables in the artificial data set. In addition, it included parameter and model selection. Consequently, the analysis of the financial indicators incurred in evaluating the performance of all models across different strategies.

In general, the first step of the forecasting window yielded better predictions than the benchmark in most approaches. However, this behavior depends on the period inside the forecasting window. As a result, the 2-steps strategy yielded more accurate results than the 1-step approach, due to the decreased dependence on uncertain predictions. However, this approach is unsuited for the final prediction of the window, in which case the h-step strategy is more reliable. In fact, this approach rendered the most accurate predictions of all strategies. Nonetheless, care should be taken on determining the forecast window size, since the greatest deviations from the true values started to occur on the fifth period.

The ARIMA model had the most detailed parameter and model selection. Thus, the forecasts of this class of models belong to one of the AR, MA, or full ARMA/ARIMA models. Furthermore, the predictions from the univariate models were more accurate than the forecasts from VAR or random forest. However, the latter model could greatly benefit from forecasting the indicators based on the fifteen shortlisted features.

In addition, the task of accurately predicting the values of financial indicators is challenging. In this sense, an alternative future direction would be to employ the h-step strategy to determine whether the indicator rises or falls. This information would provide an educated guess on the general behavior of the market towards a given indicator. In contrast, if the study requires the precise values of the indicators, then it is recommended to perform a deeper analysis on the confidence interval of the predictions.

Finally, these methodologies are able to forecast the economy in a limited range. Furthermore, the ARIMA and Random Forest models had the most accurate predictions in the h-step strategy. However, great care should be taken when applying the results of this analysis. In essence, these predictions are not enough for an automated and systematic decision, but they provide the framework for future studies.

6 REFERENCES

ALPAYDIN, E. **Introduction to Machine Learning**. Third Edition. Cambridge, MA: The MIT Press. (2014) p. 213 – 232.

BONTEMPI, G. **Machine Learning Strategies for Time Series Prediction. Summer School**. Université Libre de Bruxelles. (2013)

COGHLAN, A. **A Little Book of R For Time Series**. Wellcome Trust Sanger Institute. Available at: http://a-little-book-of-r-for-time-series.readthedocs.org/en/latest/index.html. (2015)

NAU, R. **Statistical Forecasting: notes on regression and time series analysis**. Duke University. (2015)

NG, S.; PERRON, P. **Unit Root Tests in ARMA Models with Data-Dependent Methods for the Selection of the Truncation Lag**. Journal of the American Statistical Association, 90(429), 268–281. (1995) doi.org.focus.lib.kth.se/10.2307/2291151

Pennsylvania State University. **STAT510 - Applied Time Series Analysis. Online Course**. (2015) Available at: https://onlinecourses.science.psu.edu/stat510/node/33.

TSAY, R. S. **Linear Time Series Analysis and Its Applications, in Analysis of Financial Time Series**, Third Edition, John Wiley & Sons, Inc., Hoboken, NJ, USA. (2010) doi:10.1002/9780470644560.ch2