Univariate Sensor Data Prediction Using Conventional and Machine Learning Based Time Series Techniques

Priyadarshini Mahalingam, Kalpana Dharmalingam, and Thyagarajan Thangavelu

Abstract Availability of data from sensors is becoming easy and in abundance due to the era of industrial revolution 4.0. These data carry rich information about the health condition of the process and equipments in industries along with the current status of the process from which they are acquired. Analysis of this data reveals the interaction and impact among variables involving the control loop. Forecasting and prediction of sensor data is important for the effective functioning of the predictive maintenance stream. Time stamped data can be predicted using time series forecasting techniques. In this paper, the temperature data from a temperature sensor installed to a hydraulic rig is considered for the analysis. The univariate data is predicted for future cycles using times series forecasting techniques. Comparison study between conventional and machine learning algorithms is well defined. These techniques are evaluated using different accuracy metrics like MAE, MSE and RMSE.

Keywords Time series forecasting · Univariate sensor data · One-step ahead prediction · Machine learning · Metrics

1 Introduction

Sensor data is a vital source of information in process industries. When they are acquired at regular intervals of time, they constitute the time behavioral pattern of the system and thereby time series dataset. Time series prediction is the most vast and regularly trodden field in statistical analysis. It also serves as the base work for the estimation of lifetime of sensors and in calculating their failure rate. However, the idea of which technique to be involved for prediction lies solely with the dataset at hand. Therefore, the knowledge about the dataset is important for the selection of appropriate forecast technique.

A quantum amount of work in time series analysis on univariate data sets can be seen in literatures. It shows the importance of the impact of single variable change

P. Mahalingam (✉) · K. Dharmalingam · T. Thangavelu
Department of Instrumentation Engineering, Madras Institute of Technology Campus, Anna University, Chennai, India
e-mail: darshiniipriyam@gmail.com

V. Komanapalli et al. (eds.), *Advances in Automation, Signal Processing, Instrumentation, and Control*, Lecture Notes in Electrical Engineering 700,
https://doi.org/10.1007/978-981-15-8221-9_58

in the system. Many areas conduct time series analysis investigations. Applications include electrical load forecast [1], prices and stocks, prediction of bill prices [2] etc., Time series is also applied to biological domain [3]. Many comparisons between the conventional models exists [4]. Evolving with machine learning strategies, time series analysis was investigated using supervised [5] and unsupervised algorithms. Few researchers transformed multivariate data into univariate data [6] for further study and investigations. These transformations provided suitable insights to many features. Hence, analysis for univariate datasets is important and cannot be neglected at any case.

This paper is organized such that Sect. 1 is a brief introduction to the paper. Section 2 explains the dataset that has been considered for the research work. Section 3 deals with a short note on conventional and machine learning-based time series methods like Persistent forecast model (Naïve), Simple Average model (SA), Auto Regression model (AR), Moving Average model (MA), Auto Regressive with Moving Average model (ARMA), Auto Regression Integrated Moving Average model (ARIMA), Holt Linear Trend model (HLT), Holt Winter Exponential Smoothing model (HWES), Seasonal Exponential smoothing model (SES), Linear Regression (LR), Multi Layer Perceptron (MLP), Support Vector regression (SVR), Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). Results are discussed and interpreted in Sect. 4. Section 5 deals with the conclusion and explains the future work direction for the proposed paper.

2 Dataset

In this paper, the dataset acquired for the condition monitoring of a hydraulic test rig which is available online is used. The dataset provided by [7] consists of process variables like temperature (in °C), pressure (in bar) and voluminous flow (in l/min). Here, load cycles are repeated every 60 s and the data are recorded accordingly. To illustrate the significance of time series forecasting using various techniques we only the temperature sensor reading (TS1) recorded during every cycle is considered. This data is univariate and non-stationary data. A univariate model prediction at a faster rate as proposed in [8] is the need of industries in recent times. Generally temporal datasets consists of independent time variable against any other dependent variable. In this case, time series prediction is done for the data set considering the number of cycles as a time oriented parameter.

The TS1 dataset is initially visualized to study its pattern. Figure 1 shows the temperature dataset which is considered for the analysis. It can be seen that the temperature is increasing in the initial cycles and tends to maintain a value and then drop during the last phase. This pattern remains throughout the entire cycle. The size of the dataset is 2204 $\times$ 60. A zoom plot of the dataset is shown in Fig. 2.

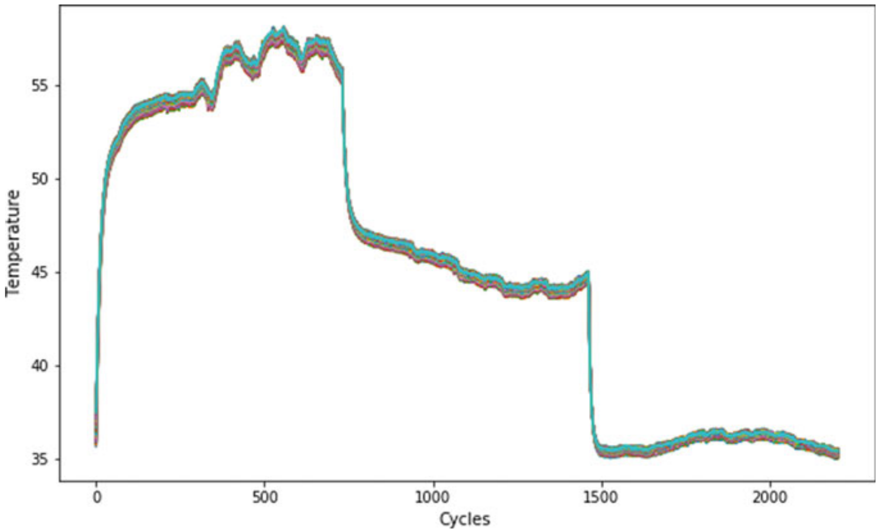

Fig. 1 Visualization of the temperature dataset

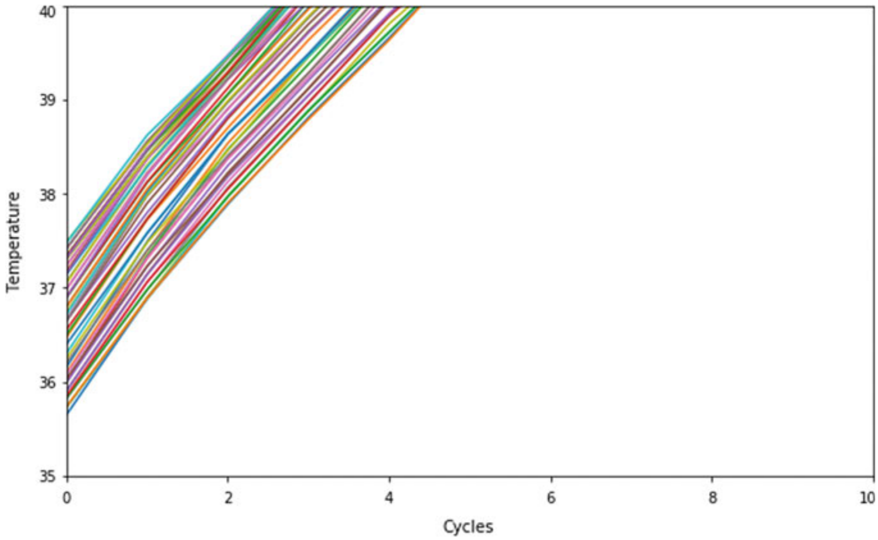

Fig. 2 Zoom out view of Fig. 1

3 Time Series Techniques

The difference between forecasting and prediction is very narrow. The former expects a range of values in the future and the latter is an estimate of the future value with regards to the previous value. The general procedure in time series forecasting involves (1) Visualizing the data to study its structures. (2) Converting non-stationary data into stationary data. (3) Deciding the appropriate values for building the model. (4) Developing time series model (5) Prediction (6) Evaluation using metrics.

3.1 Conventional Time Series Prediction Techniques

For simplicity in explanation, only 2204 $\times$ 1 size of the temperature dataset is considered. On visualization, null hypothesis is detected. To achieve the hypothesis, static tests and preprocessing techniques are performed. The static test results for the non-stationary data are as shown in Table 1.

The static test results for the stationarized data are shown in Table 2. From both tables, it can be seen that the necessary conditions for Stationarity like (i) Critical value > test value and (ii) p value < 0.5 is satisfied. Thus the hypothesis is achieved. A selection matrix as shown in Fig. 3 is developed to ease the selection of suitable model based on few parameters.

Figure 3 shows the selection matrix of various conventional time series prediction techniques and the most suitable type of situation in which they fit in for better performance. Here, the colored portion indicates that the method is capable of handling the condition mentioned. The following are the conditions against which each method is mapped.

1—Univariate dataset
2—Multivariate dataset
3—One-step ahead prediction
4—Multistep ahead prediction
5—Trend
6—Seasonality
7—Exogenous input

Table 1 Static test result for the non-stationary data

Result	Test static	p-value	No. of. lags used	No. of. observations used	Critical value (1%)	Critical value (5%)	Critical value (10%)
Values	−0.558	0.8801	6	2198	− 3.433	− 2.862	− 2.567

Table 2 Static test result for the stationary data

Result	Test static	p-value	No. of. lags used	No. of. observations used	Critical value (1%)	Critical value (5%)	Critical value (10%)
Values	− 3.032	0.0320	6	2198	− 3.433	− 2.862	− 2.567

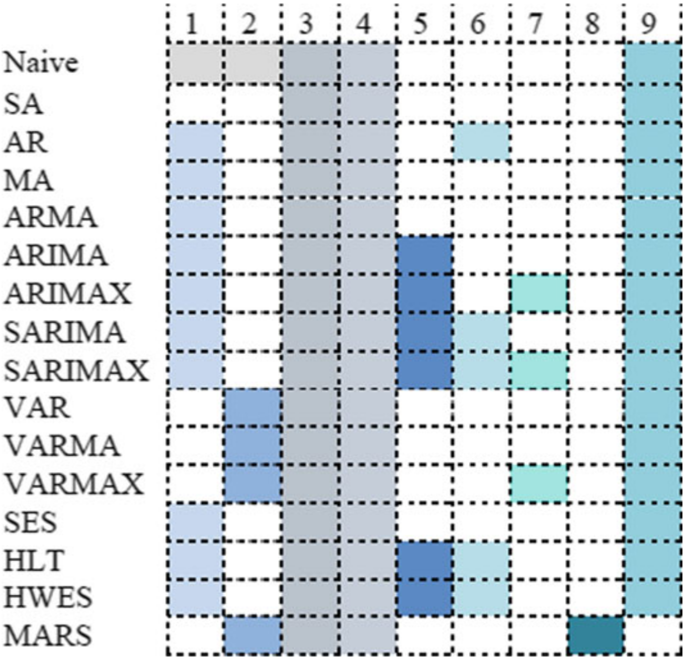

Fig. 3 Selection matrix for conventional time series method

Table 2 Basic forecast metrics

Metric	Pros	Cons
AE—absolute error	Easy (no squares or square roots required	Error can be due to round off values leading to inexact result
ME—mean error	Easy and straight forward	Opposite quantity errors cancel out
MAD—mean absolute deviation	Easy (no squares or square roots required)	Not accurate for randomly distributed data
MAE—mean absolute error	Simpler to interpret. An average of absolute errors	Not used for evaluating different series
MSE—mean squared error	Easy. Converts all errors into a positive quantity	Not applicable for models having too high and too low errors
RMSE—root mean squared error	Overcomes the disadvantage of MSE	Quite hard to interpret when dataset is not consistent in nature
MAPE—mean absolute percentage error	Can be used for evaluating two or more different series	Gives indeterminant values when errors are zero
MRAE—mean absolute relative error	Suitable for different time series prediction	Not applicable for data value zero

8—Parametric method
9—Non-parametric method.

However, a vast amount of research is being done in mapping the un-colored portions. Hence the illustration here is a most general selection criterion for the conventional methods. Persistent model (Naive method) is the most baseline model and can be used as a simple tool for forecasting any kind of dataset. Models like AR, MA, ARMA, ARIMA, ARIMAX, SARIMA, SARIMAX, SES, HLT and HWES work well for univariate data analysis rather than multivariate data analysis. Among them ARIMAX and SARIMAX require an exogenous input and hence, these models are not considered.

From Fig. 1, it can be concluded that the dataset consists of a constant trend and an additive seasonality in it. Hence, only conventional models as presented in Table 2 are developed and their metrics are calculated.

3.2 Machine Learning-Based Time Series Prediction Techniques

For simplicity in explanation, only 2204 $\times$ 1 size of the temperature dataset is considered. On visualization, null hypothesis is detected. To achieve the hypothesis, static tests and preprocessing techniques are performed. The static test results for the non-stationary data are as shown in Table 1. Though conventional methods are easy and work very similar to supervised algorithms, few drawbacks sets conventional methods at a lag. To discuss a few, a large dataset is a challenge to any conventional model to retrain every time to forecast in a new horizon. This is because, in conventional modeling, training is done for each prediction and the model tends to change with ease. Also, conventional models cannot work with the hidden patterns in the data. Hence, to avoid such lags, machine learning models are required. The various machine learning-based time series models for sales time series prediction are discussed in [9]. Forecasting of the energy consumption time series using machine learning-based techniques are performed in [10].

In this paper, more recent and common machine learning algorithms are applied for the univariate sensor datasets and their results are discussed. A selection matrix need not be developed since machine learning algorithms can be made flexible considering only the shortcomings of the algorithm value selection.

3.3 Forecast Metrics for Evaluation

Evaluation is necessary for any technique to compare their performance strengths. In statistics, these evaluation criteria are called metrics. Forecast metrics are empirical formulas that are utilized to evaluate the performance of a particular model. There are many accuracy metrics available in literature. The accuracy metrics are classified into scale dependent metrics, percentage error metrics, relative error metrics and scale free metric in [11]. It was concluded that the metrics must be classified as Primary metrics, extended metrics, Composite metrics and Hybrid metrics in [12]. Table 2 lists few commonly used accuracy metrics in time series prediction with their advantages and disadvantages.

4 Results and Discussions

The forecast was done with 80–20 training and test set. The training and test prediction is shown in Fig. 4. The forecast done through all the above discussed method in Sect. 3 proved 95% confidence interval as shown in Fig. 5.

From the forecast models built, their performance using few fundamental accuracy metrics is investigated.

Table 3 lists accuracy metrics for the conventional models built for time series prediction. It can be seen that, though all the above models support univariate dataset, the error indicated is not the same. This effect depends upon the structure of the dataset. For a detailed explanation, Simple averaging (SA) model simply captures the average between the previous points and projects it into the future. Hence, it cannot predict any trend in the dataset. Therefore, large squared and absolute error has occurred. In the case of AR, MA and ARMA, the data is non-stationary and future points are predicted only based on a linear function of past data and residual

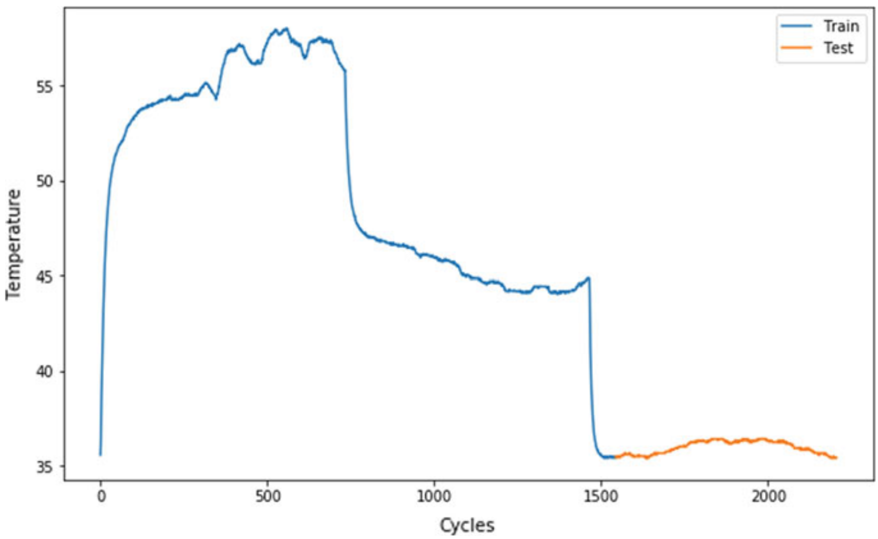

Fig. 4 Training and testing validation

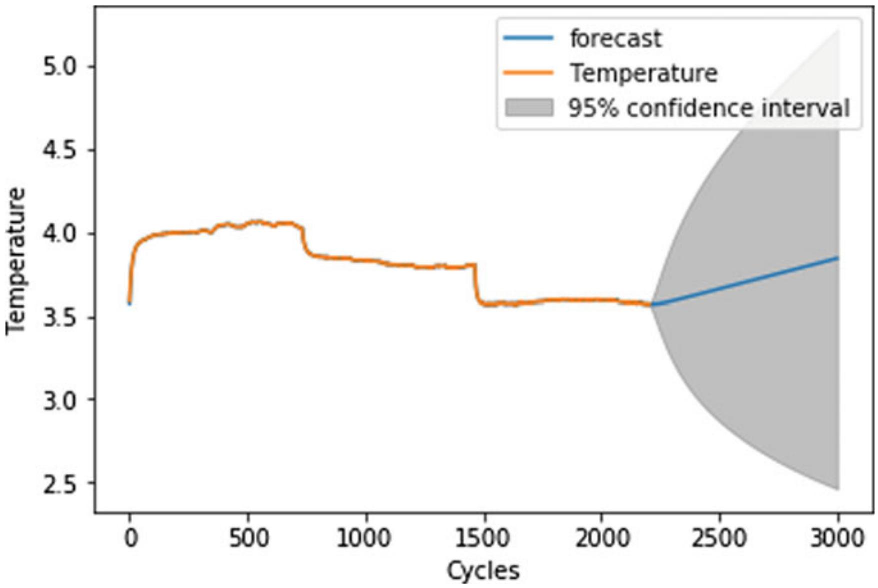

Fig. 5 Forecast with 95% confidence interval

Table 3 Accuracy metrics for conventional time series prediction models

	Naïve	SA	AR	MA	ARMA
MAE	0.56176	13.6269	4.5825	7.9965	4.09158
MSE	0.4313	185.8107	27.1382	64.0601	23.7744
RMSE	0.6568	13.63129	5.20994	8.00376	4.8759

	ARIMA	HLT	HWES	SES
MAE	1.9192	1.22359	26.192	0.5593
MSE	5.3638	1.83033	2583.190	0.42697
RMSE	2.316	1.3529	50.8251	0.65343

errors. No accountability to stationarize the data is performed. In ARIMA model, the factor 'I' accounts for the stationarity of the data through differencing. Hence, the RMSE is reduced when compared to previous models.

HLT method does not capture the additive seasonality in the dataset and hence fails. Comparing HWES and SES, the HWES method applies 'triple exponential smoothing' and hence the model has over-fitted. However, SES performs 'single exponential smoothing'. Hence, an acceptable value of RMSE is obtained. It can be noted that, the value given by SES model is already predicted through a simple baseline model called the 'Naïve' or 'Persistent model'.

There is a rising requirement to compare conventional and machine learning models beforehand like in [13]. Table 4 presents the comparison of RMSE metric

Table 4 Comparison of RMSE metric for conventional and machine learning models

	Naïve	SA	AR	MA	ARMA
RMSE	0.6568	13.63129	5.20994	8.00376	4.8759

	ARIMA	HLT	HW	SES
RMSE	2.316	1.3529	50.8251	0.65343

	LR	MLP	SVR	CNN	RNN
RMSE	0.6733	2.340	11.093	34.130	43.727

for conventional and machine learning models built. In Table 4, the RMSE of linear regression algorithm is acceptable since it is the most common supervised algorithm working similar to exponential trend methods. It can be seen that ARIMA and MLP returns similar error values. The reason for this lies in the execution structure of both though they are mostly for linear and non-linear components respectively. SVR failed because, a hyper-plane with minimum margin could not be developed since the data has gradually increasing and decreasing patterns. CNN and RNN prove good results in [14] due to their massive dataset and proper selection of pre-processing techniques. However, there is a peak rise in RMSE values for CNN and RNN can be seen because they tend to work well with larger datasets and features. When it is used for smaller datasets, they tend to over-fit.

From Table 4, it can be concluded that, for any time series forecast one cannot directly apply the machine learning techniques. A thorough study of dataset is required to identify the appropriate model to be built.

This also facilitates the development of hybrid structure of model building proposed in [15].

5 Conclusion and Future Work

This paper discussed the application of both conventional and machine learning-based time series forecast techniques. From the analysis, it is concluded that, the appropriate model fitness depends on the dataset that is considered. The conventional and machine learning-based models work uniquely on the same dataset. Also, the algorithm with highest accuracy and least error has to be selected as the fittest model and included in our further studies. This will eliminate the propagation of errors in consecutive procedures. In our future work, the forecast output will be taken as a factor for the prediction of the degradation of sensors and thereby their lifetime.

Acknowledgements This work is supported by the Centre for Research, Anna University, Chennai, India by providing financial assistance through Anna Centenary Research Fellowship (ACRF).

References

1. Taylor JW, McSharry PE (2007) Short-term load forecasting methods: an evaluation based on European data. IEEE Trans Power Syst 22(4):2213–2219
2. Zareipour H, Cañizares CA, Bhattacharya K, Thomson J (2006) Application of public-domain market information to forecast Ontario's wholesale electricity prices. IEEE Trans Power Syst 21(4):1707–1717
3. Hoyer D, Kaplan D, Schaaff F, Eiselt M (1998) Determinism in bivariate cardiorespiratory phase-space sets. IEEE Eng Med Biol Mag 17(6):26–31
4. Chatterjee S, Bard JF (1985) A comparison of Box-Jenkins time series models with autoregressive processes. IEEE Trans Syst Man Cybern 2:252–259
5. German G, Paz Sesmero M, Sanchis A (2016) Forecasting time series by an ensemble of artificial neural networks based on transforming the time series. In: 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC), IEEE, pp 004769–004774
6. Li J, Pedrycz W, Jamal I (2017) Multivariate time series anomaly detection: a framework of hidden markov models. Appl Soft Comput 60:229–240
7. Helwig N, Pignanelli E, Schütze A (2015) Condition monitoring of a complex hydraulic system using multivariate statistics. In: IEEE International Instrumentation and Measurement Technology Conference (I2MTC) Proceedings, pp 210–215)
8. Majidpour M, Nazaripouya H, Chu P, Pota HR, Gadh R (2019) Fast univariate time series prediction of solar power for real-time control of energy storage system. Forecasting 1(1):107–120
9. Pavlyshenko BM (2019) Machine-learning models for sales time series forecasting. Data 4.1, pp 15
10. Chou J-S, Tran D-S (2018) Forecasting energy consumption time series using machine learning techniques based on usage patterns of residential householders. Energy 165:709–726
11. Hyndman RJ, Koehler AB (2006) Another look at measures of forecast accuracy. Int J Forecast 22(4):679–688
12. Botchkarev A (2018) Performance metrics (error measures) in machine learning regression, forecasting and prognostics: properties and typology. arXiv preprint arXiv:1809.03006
13. Wei N, Li C, Peng X, Zeng F, Lu X (2019) Conventional models and artificial intelligence-based models for energy consumption forecasting: A review. J Petrol Sci Eng pp 106187
14. Fang X, Yuan Z (2019) Performance enhancing techniques for deep learning models in time series forecasting. Eng Appl Artif Intell 85:533–542
15. Büyükşahin ÜÇ, Ertekin Ş (2019) Improving forecasting accuracy of time series data using a new ARIMA-ANN hybrid method and empirical mode decomposition. Neurocomputing 361:151–163