International Journal of Hybrid Information Technology
Vol. 10, No. 6 (2017), pp.13-22
http//dx.doi.org/10.14257/ijhit.2017.10.6.02

# Annual Automobile Sales Prediction Using ARIMA Model

Sana Prasanth Shakti, Mohan Kamal Hassan, Yang Zhenning,
Ronnie D. Caytiles[#] and Iyengar N.Ch.S.N.

*School of Computer Science and Engineering. VIT University, Vellore, T.N., India*
*[*]School of Information Technology and Engineering,*
*VIT University, Vellore, T.N. India.*
*[#]Multimedia Engineering department, Hannam University, Daejeon, Korea*
*[1]prasanth.shakti@gmail.com, [2]mohan.kamalhassan@gmail.com*
*yang.zhenning2015@vit.ac.in, nchsniyengar48@gmail.com,*
*rdcaytiles@gmail.com*

***Abstract***

*Sales forecasting is a most important application in industries and has been one of the most scientifically and technologically challenging problems around the world. One approach of prediction is to spot patterns in the past, when it is known in advance what followed them and verify it on more recent data. If a pattern is followed by the same outcome frequently enough, it can be concluded that it is a genuine relationship. Because this approach does not assume any special knowledge or form of the regularities, the method is quite general applicable to other series not just climate. Sales prediction phenomena have many parameters like Number of sales, production, Consumed cost and Time required that are impossible to enumerate and measure. In this paper, we are going to use the ARIMA model for predicting the number of sales for a Time series data. The dataset tractor sales data for a period of ten years (2003-2014) obtained from the Mahindra Tractors Company are used from which use to classify the performance by drawing various scattered plots and graphs. The result of the ARIMA results shows that which predicts better for the sales prediction of the next following 5 years.*

***Keywords****: Sales Prediction, ARIMA, ARMA, Time Series*

## 1. Introduction

Sales prediction is a complex process and a challenging task for researchers. It includes expertise in multiple disciplines. The prediction of atmospheric parameters is essential for various applications. Accurate prediction of Sales parameters is a difficult task due to the dynamic nature of atmosphere. Generally prediction is done on Time series data. A time series is a sequence of observed values of some entity that is measured at different points in time. With the advancement of collecting data, huge amounts of data have been collected making it impossible to be processed manually. This is where the time series analysis has to be automated and take advantage of modern computing mechanisms.

Various techniques like linear regression, auto regression, Multi-Layer Perceptron, Radial Basis Function networks are applied to predict atmospheric parameters like temperature, wind speed, rainfall, meteorological pollution *etc.* It was found that the nonlinear operator equations governing the atmospheric system are the ones who can better understand the dynamics of atmosphere. In the recent past many forecast methods have been developed using Artificial Neural Networks (ANNs). Neural network techniques have the potential to handle complex, nonlinear problems in a better way when compared to traditional techniques. However systems developed using neural network model suffer from certain drawbacks like local minima, model over fitting *etc.*

ISSN: 1738-9968 IJHIT

In this paper we are going use ARIMA model for the predicting the automobile sales. The performance of ARIMA is compared with SVM and show that which is better for predicting the temperature with a short time. We have collected data set from Jan-2003 to Dec-2014 from the Mahindra tractor sales. Here, we have taken "Number of tractors sold" parameter where we are going to apply ARIMA to predict the sales for the next following 5 years.

## 2. Literature Survey

The first operational numerical sales prediction model consisted of only one layer and therefore it could model only the slight variation of the mean vertical structure of the sales. Computers now permit the development of multilevel (usually about 10–20) models that could resolve the vertical variation of the production, consumption, time. These multilevel models predict the fundamental sales variables for large scales of motion.

M. M. Elkateb *et al.* (1998) "A comparative study of medium-weather-dependent load forecasting using enhanced artificial/fuzzy neural network and statistical techniques". Monthly peak load demand of Jeddah area for the past nine years was used for investigation. The first seven years data was used for training while the prediction was carried out for the following two years. First, Minitab statistical software package was used for peak load prediction using Autoregressive Integrated Moving Average (ARIMA) technique, and an average error value of 11.7% is achieved. Next, an Artificial Neural Network (ANN) was utilized and several suggestions are implemented to build an adaptive form of ANN. Direct ANN implementation shown poor performance. Also, Fuzzy Neural Network (FNN) was also examined but showed comparatively better performance. The modeling of the trend of peak load demand is incorporated by introducing "time index feature" and that clearly enhanced the performance of both ANN (6.8% error) and FNN (4.7% error).

Raymond Y.C. Tse, (1997) suggested that the following two questions must be answered to identify the data series in a time series analysis: (1) whether the data are random; and (2) are there any patterns. This is followed by another three steps of model identification, parameter estimation and testing for model validity. If a series is random, the correlation between successive values in a time series is close to zero. If the observations of time series are statistically dependent on each another, then the ARIMA is appropriate for the time series analysis.

Meyler *et al* (1998) drew a framework for ARIMA time series models for forecasting Irish inflation. In their research, they emphasized heavily on optimizing forecast performance while focusing more on minimizing out-of-sample forecast errors rather than maximizing in-sample 'goodness of fit'.

In 2000 Toth *et al*. compared short-term rainfall prediction models for real-time flood forecasting. They applied three time series models, auto-regressive moving average (ARMA), ANN and k-nearest-neighbors (KNN) method for forecasting storm rainfall occurring in the Sieve River basin, Italy, in the period 1992- 1996 with lead times varying from 1 to 6 h. The result showed that the ANN performed the best in the improvement of the runoff forecasting accuracy when the predicted rainfall was used as inputs of the rainfall run-off model.

Contreras *et al* (2003) in their study, using ARIMA methodology, provided a method to predict next-day electricity prices both for spot markets and long-term contracts for mainland Spain and Californian markets.

In, "A comparative analysis for wind speed Prediction", authors Tarade R.S, Katti P.K presented a wind speed predictor based on Autoregressive Integrated Moving Average (ARIMA) which is able to predict short term wind speed, which is essential in order to prevent systems inaction from the effects of strong winds. It also helps in using wind energy as an alternative source of energy, mainly for electrical power generation. Wind

speed prediction has applications in military and civilian fields for air traffic control, rocket launching, ship navigation *etc*. The mean squared error (MSE) was 3.57.

## 3. Methodology

### 3.1. ARIMA: (Auto Regressive Integrated Moving Average)

*ARIMA model is done on time series data.* Time series data is a sequence of observations collected from a process with *equally* spaced periods of time.

Examples: Industrial Averages, Daily data on sales, Daily Customers.

Stages of Time series model process using ARIMA:

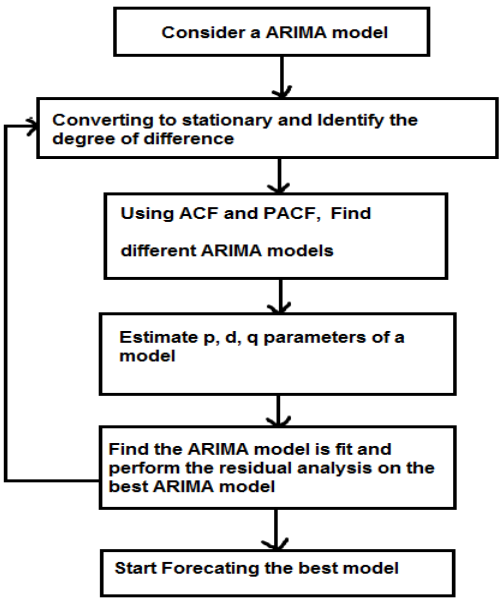



**Figure 1. Flow Diagram of ARIMA Model**

ARIMA is also known as Box-Jenkins approach. To build a time series model issuing ARIMA, we need to study the time series and identify p,d,q. Where,

p – *A*uto *R*egressive (Auto Correlation)

d - *I*ntegrated (Stationary / Trend)

q - *M*oving *A*verage (Shocks / Error)

- **Identification:** Determine the appropriate values of p, d, & q using the ACF, PACF, and unit root tests. p is the AR order, d is the integration order, q is the MA order
- **Estimation:** Estimate an ARIMA model using values of p, d, & q you think are appropriate.
- **Diagnostic checking:** Check residuals of estimated ARIMA model(s) to see if they are white noise; pick best model with well-behaved residuals.

**Forecasting:** Produce out of sample forecasts or set aside last few data points for in-sample forecasting.

### 3.2. Autoregressive (AR) Process:

Series of current values depend on its own previous values.

In an AR(p) model the future value of a variable is assumed to be a linear combination of p past observations and a random error together with a constant term.

Mathematically the AR(p) model can be expressed as :

$$y_t = c + \sum_{i=1}^{p} \varphi_i y_{t-i} + \varepsilon_t = c + \varphi_1 y_{t-1} + \varphi_2 y_{t-2} + \ldots\ldots\ldots\ldots + \varphi_p y_{t-p} + \varepsilon_t \quad (1)$$

Here $y_t$ and $\varepsilon_t$ are respectively the actual value and random error(random shocks) at time period t, $\phi_i$ (i = 1,2,..., p) are model parameters and c is a constant. The integer constant p is known as the order of the model.

### 3.3 Moving Average (MA) Process:

The current deviation from mean depends on previous deviations.

An AR(p) model regress against past values of the series, an MA(q) model uses past errors as the explanatory variables.

The MA(q) model is given by :

$$y_t = \mu + \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} + \varepsilon_t = \mu + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \ldots\ldots\ldots\ldots + \theta_q \varepsilon_{t-q} + \varepsilon_t \quad (2)$$

Here μ is the mean of the series, ( j=1,2,...,q) $\theta_j$ = are the model parameters and q is the order of the model.

Conceptually a moving average model is a linear regression of the current observation of the time series against the random shocks of one or more prior observations.

### 3.4. Stationary Series (Integrated):

In order to model a time series with the Box-Jenkins approach, the series has to be stationary. In practical terms, the series is stationary if tends to wonder more or less uniformly about some fixed level. In statistical terms, a stationary process is assumed to be in a particular state of statistical equilibrium, *i.e.*, p(xt) is the same for all t.

In order to achieve a series as stationary we need to do Regular differencing (RD),

$$\text{(1st order)} \ \nabla x_t = (1 - B)x_t = x_t - x_t\text{-}1 \quad (3)$$

$$\text{(2nd order)} \ \nabla^2 x_t = (1 - B)^2 x_t = x_t - 2x_{t-1} + x_{t-2} \quad (4)$$

"B" is the backward shift operator

It is unlikely that more than two regular differencing would ever be needed

## 4. Dataset and Preprocessing

The tractor sales data for a period of ten years (2003-2014) obtained from the Mahindra tractors company, from which data is used to build the models and the prediction for the next following five years between January and December is obtained by the test models. The database includes readings of several weather parameters recorded at every half hour interval. The yearly tractor sales is extracted from this database and used for this work. The real world databases are highly susceptible to noisy and missing data. The data can be preprocessed to improve the quality of data and thereby improve the prediction results. In this work, the data taken had applied cleaning and transformation. Data cleaning fills in the missing values, while data transformation improves the accuracy, speed and efficiency of the algorithms used.

The missing value for number of sales in a day is replaced with the mean of number of sales for that month while building the SVM model. The data is normalized using Z-score normalization where the values of an attribute A, are normalized based on the mean ($\bar{A}$) and standard deviation ($\sigma A$) of A. The normalized value $V'$ of $V$ can be obtained as,

$$V' = (V - \bar{A}) / \sigma_A \quad (5)$$

However, it was found that the negative values generated by Z-score normalization cannot be fed as input to the ARIMA model.

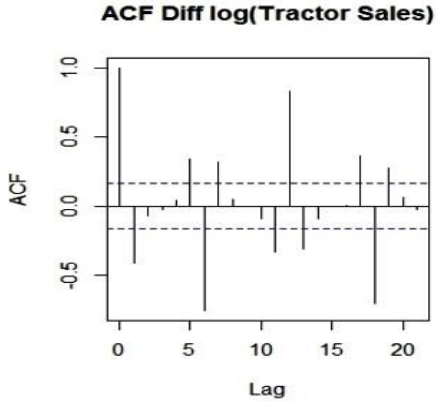
## 5. Results and Discussion

The performance of ARIMA is checked by plotting the graph for the confusion matrix, which is generated for the temperature against years. A confusion matrix can help visualize the results of a ARIMA classification algorithm. In this paper the implementation is done in the programming language R. Here, we can show that results by compare to other models ARIMA can perform better.

### ARIMA Model

Here, the data is predicted from the taken dataset by first converting the data into stationary. To make stationary we have to find the difference on mean of Number of sales. Then the converted data is log transformed on variance such that the final data should of log transform value for the both mean and variance.

Now plot the ACF and PACF graphs to identify the potential AR and MA model.



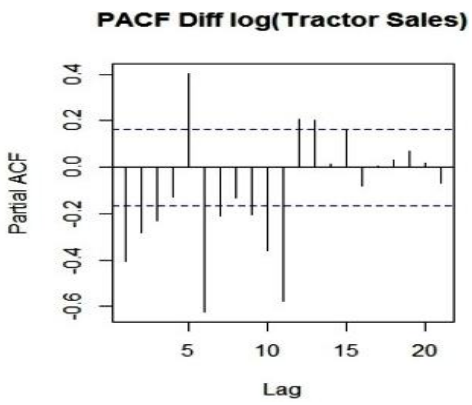




**Figure 2. ACF Graph for AR Model** **Figure 3. PACF Graph for the MA Model**

From the two ACF and PACF graphs we have to find the potential ARIMA fit from which we are going to predict the sales forecasting. When choosing the model which is best fit it should be based on AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion). Normally, the model which has lowest AIC value is taken is best fit.

```
Series: log10(data)
ARIMA(0,1,1)(0,1,1)[12]

Coefficients:
          ma1     sma1
      -0.4047  -0.5529
s.e.   0.0885   0.0734

sigma^2 estimated as 0.0002571:  log likelihood=354.4
AIC=-702.79   AICc=-702.6   BIC=-694.17

Training set error measures:
                      ME       RMSE        MAE         MPE      MAPE      MASE
Training set 0.0002410698 0.01517695 0.01135312 0.008335713 0.4462212 0.2158968
                    ACF1
Training set 0.01062604
```

**Figure 4. Choosing Best ARIMA Fit**

As expected our model has I (or integrated) component equal to 1. This represents differencing of order 1. There is additional differencing of lag 12 in the above best fit model. Moreover, the best fit model has MA value of order 1. Also, there is seasonal MA with lag 12 of order 1.

The final graph is plotted for the best fit ARIMA model of number of sales next following years. The following is the output with forecasted values of tractor sales in blue. Also, the range of expected error (*i.e.* 2 times standard deviation) is displayed with orange lines on either side of predicted blue line.

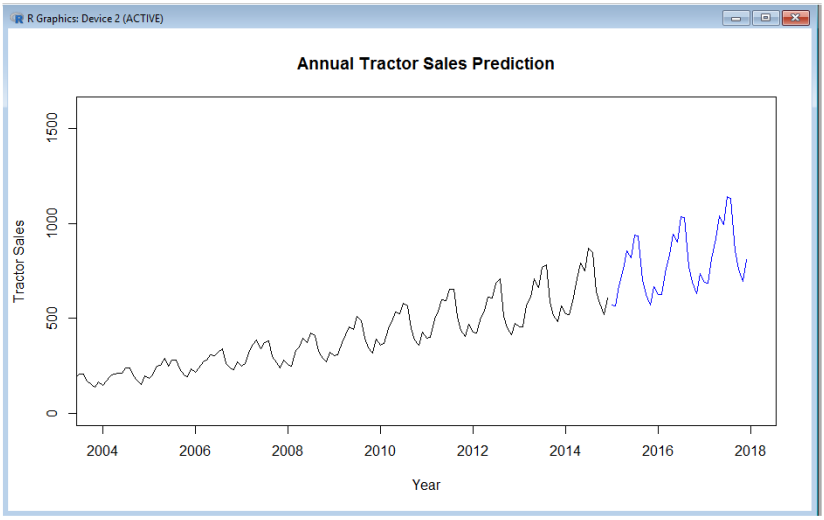



**Figure 5. ARIMA Model Prediction up to 2018**

Here forecast for a long period like 2 years is an ambitious task. The major assumption here is that the underlining patterns in the time series will continue to stay the same as predicted in the model. A short term forecasting model, say a couple of business quarters or a year, is usually a good idea to forecast with reasonable accuracy. A long term model like the one above needs to evaluate on a regular interval of time (say 6 months). The idea is to incorporate the new information available with the passage of time in the model.

The prediction for 2 years is shown in month-wise, this can sometimes vary with actual result.

After predicting the number of sales, ensure that there are no more information is left for prediction that is there are no residuals in the ARIMA model. We can find if there any residuals bycreate an ACF and PACF plot of the residuals of our best fit ARIMA model *i.e.* ARIMA(0,1,1)(0,1,1)[12].

```
$pred
          Jan      Feb      Mar      Apr      May      Jun      Jul      Aug
2015 2.754168 2.753182 2.826608 2.880192 2.932447 2.912372 2.972538 2.970585
2016 2.796051 2.795065 2.868491 2.922075 2.974330 2.954255 3.014421 3.012468
2017 2.837934 2.836948 2.910374 2.963958 3.016213 2.996138 3.056304 3.054351
          Sep      Oct      Nov      Dec
2015 2.847264 2.797259 2.757395 2.825125
2016 2.889147 2.839142 2.799278 2.867008
2017 2.931030 2.881025 2.841161 2.908891

$se
            Jan         Feb         Mar         Apr         May         Jun         Jul
2015 0.01603508 0.01866159 0.02096153 0.02303295 0.02493287 0.02669792 0.02835330
2016 0.03923008 0.04159145 0.04382576 0.04595157 0.04798329 0.04993241 0.05180825
2017 0.06386474 0.06637555 0.06879478 0.07113179 0.07339441 0.07558934 0.07772231
            Aug         Sep         Oct         Nov         Dec
2015 0.02991723 0.03140337 0.03282229 0.03418236 0.03549035
2016 0.05361850 0.05536960 0.05706700 0.05871534 0.06031866
2017 0.07979828 0.08182160 0.08379608 0.08572510 0.08761165
```

**Figure 6. Prediction for Year 2017 & 2018 in Month-wise**

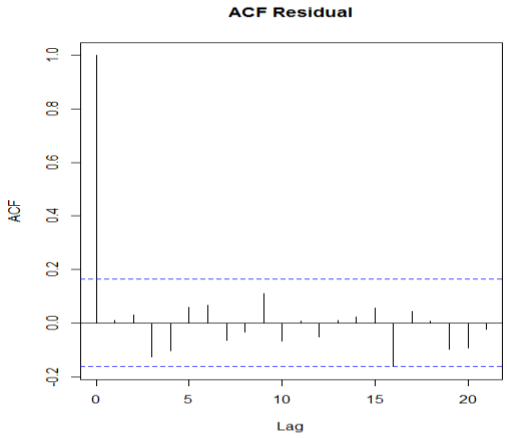



**Figure 7. ACF Residual Graph for the AR Model**

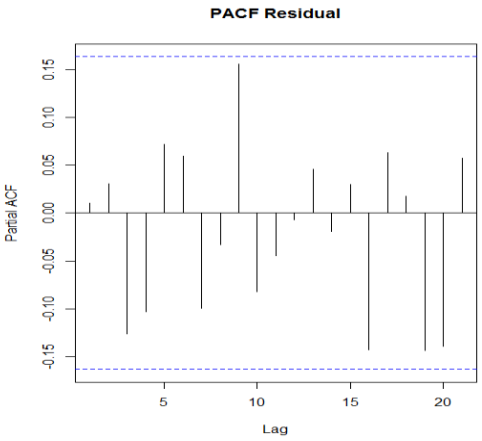



**Figure 8. PACF Residual Graph for the MA Model**

Since there are no spikes outside the insignificant zone for both ACF and PACF plots we can conclude that residuals are random with no information in them.

## 6. Comparison

The best values of parameters in the ARIMA models have been shown in Figure 4. According to Figure 4, it was determined that by increasing in number of autoregressive and the moving average parameters, error rate was reduced. Thus, as was clear from Figure 4, the best ARMA model has three seasonal autoregressive parameters and four seasonal moving average parameters. The best ARIMA model had zero autoregressive parameters, one moving average parameter, zero seasonal autoregressive parameter, and one seasonal moving average parameter, and $d = 1$. From this we can know that with increasing number of parameters has effect on performance improvement of ARMA and ARIMA models.

The best structure between ARMA, ARIMA can be decided by knowing the benchmark index RMSE (Root mean square error) value, it is calculated mathematically by,

$$RMSE = \sqrt{\sum_{i=1}^{n} (Q_{ci} - Q_{oi})^2 / n} \tag{6}$$

Where $Q_{ci}$ is the computational record in month $i$, the $Q_{oi}$ observational record in month $i$, and $n$ is the number of data.

From the previous research we know that the benchmark index RMSE for ARMA(1, 0)(2,1)$_{12}$ forecasting model was equal to 0.1732 and the benchmark error index RMSE for our work in the forecasting model ARIMA(0, 1, 1)(0, 1, 1)$_{12}$ data was equal to 0.01517 and it was chosen as the best model to forecast automobile sales, from the all models between ARMA and ARIMA.

The ARIMA model has a better performance than ARMA model because it makes time series stationary, in both training and forecasting phases.

## 7. Conclusion

In this paper the performance of ARIMA is done by showing different visualization graphs. Results obtained shows that ARIMA performs better for the next following years. Number of sales parameter observed that it has significant effect for ARIMA model, that is it predicts has a wide deviation from the data taken from the previous years. That is ARIMA model predicted values has a large difference from the expected values. As we

know there no model which is predicted perfectly when it comes to sales forecasting. But as of now ARIMA model is more suitable for sales forecast for the static time series data.

## References

[1] D. Riordan and B. K. Hansen, "A fuzzy case-based system for weather prediction", Engineering Intelligent Systems, vol. 10, no. 3, **(2002)**, pp. 139-146.
[2] P. Guhathakurtha, "Long-Range monsoon rainfall prediction of 2005 for the districts and sub-division Kerala with artificial neural network", Current Science, vol. 90, no. 6, **(2006)**, pp. 773-779.
[3] H. H. Brian, "Insights into neural network forecasting of time sense corresponding to ARMA (p, q) structure", Omega, vol. 299, no. 3, **(2001)**, pp. 273-289.
[4] A. Wadia and M. T. S. Ismail, "Selecting Wavelet Transforms Model in Forecasting Financial Time Series Data Based on ARIMA Model", Applied Mathematical Sciences, vol. 5, no. 7, **(2011)**, pp. 315 – 326.
[5] C. Charisios, C. Michalakelis and D. Varoutas, "Forecasting with limited data: Combining ARIMA and diffusion models", Technological forecasting and social change, vol. 77, no. 4, **(2010)**, pp. 558-565.
[6] N. R. Pal, S. Pal, J. Das, and K. Majumdar, "SOFM-MLP: A Hybrid Neural Network for Atmospheric Temperature Prediction", IEEE Transactions on Geoscience and Remote Sensing, vol. 41, no. 12, **(2003)**, pp. 2783-2791.
[7] G. P. Zhang, "Time series forecasting using a hybrid ARIMA and neural network model", Neurocomputing, vol. 50, **(2003)**, pp. 159-175.
[8] Thiesing and Vornberger, "Sales forecasting using neural networks", Proc. of the International Conference on Computational Intelligence, Theory and Applications, vol. 4, **(1997)**, pp. 321-328.
[9] J. Wang, J. Liang, J. Che and D. Sun "ARMA Model identification using Particle Swarm Optimization Algorithm", Int. conference on Computer Science and Information Technology, **(2008)**, pp. 223-227.
[10] A. P. Ansuj, "Sales forecasting using time series and neural networks", Computers & Industrial Engineering, vol. 31, no. 1-2, **(1996)**, pp. 421-424.
[11] C. Chris, "The analysis of time series: an introduction", CRC press, **(2013)**.

## Authors

**Mohan Kamal Hassan**, received the B.Tech (Information Technology) degree from the Mahatma Gandhi Institute of Technology, Hyderabad in 2016 which is Affiliated to JNTUH University, Hyderabad. Currently he is pursuing M. Tech (Bigdata Analytics) degree from Vellore Institute of Technology, Vellore. His area of interest in Data Analytics and its related Big Data Technology.

**Sana Prasanth Shakti**, received the B.Tech (Computer Science and Engineering) degree from the Saveetha School of Engineering, Saveetha University, Chennai in 2015. Currently he is pursuing M. Tech (Bigdata Analytics) degree from Vellore Institute of Technology, Vellore. His area of interest in Data analytics and its behaviour on various fields.

**Yang Zhenning**, he is pursuing M.ScComputer Science at School of Computing Science and Engineering, VIT University, Vellore. His area of interests are Algorithm design an Pattern Recognition, operating Systems and cloud computing

**N. Ch. S. N. Iyengar (b 1961)**, he currently Senior Professor at the School of Computer Science and Engineering at VIT University, Vellore-632014, Tamil Nadu, India. His research interests include Agent-Based Distributed Computing, Intelligent Computing, Network Security, Secured Cloud Computing and Fluid Mechanics. He had 30+ years of experience in teaching and research, guided many scholars, has authored several textbooks and had nearly 200+ research publications in reputed peer reviewed international journals. He served as PCM/reviewer/keynotespeaker/Invited speaker for many conferences. He serves as editorial board member for many international journals, reviews papers for many conferences with an interest of serving to the education community

**Ronnie D. Caytiles**, he had his Bachelor of Science in Computer Engineering- Western Institute of Technology, Iloilo City, Philippines, and Master of Science in Computer Science– Central Philippine University, Iloilo City, Philippines. He finished his Ph.D. in Multimedia Engineering, Hannam University, Daejeon, Korea. Currently, he serves as an Assistant Professor at Multimedia Engineering department, Hannam University, Daejeon, Korea. His research interests include Mobile Computing, Multimedia Communication, Information Technology Security, Ubiquitous Computing, Control and Automation.