Analysis of crucial oil gas and liquid sensor statistics and production forecasting using IIOT and Autoregressive models

Anurag Kumar Singh[1], R.K. Pateriya[2]

[1]M. tech Student, Dept. of Computer Science and Engineering, MANIT, Bhopal, India
[2]Associate Professor, Dept. of Computer Science and Engineering, MANIT, Bhopal, India

***

Abstract - *IIOT devices and analytics based on them has been a breakthrough in today's engineering. IIOT is being used in almost every fields of technology. In Oil and Gas industry, IIOT based analytics is of crucial importance. Until recently, the oil field manager had to manually visit the wells and check each of the sensors and equipments for collecting real time pro-duction data that have been achieved over the past times. Also, no hint about wear and tear of the equipments can be known prior to the actual failure of the concerned machines. This paper presents our approach where we apply auto-regressive model (ARIMA) to forecast the sensor statistics like production values that might be seen in future using time series data. This paper focuses how much advancement in IIOT technologies have helped us to analyze past data to fore-cast future data beforehand with good accuracy for oil, gas or liquid generation.*

***Key Words*: ARIMA, Oil & Gas production, Dickey-Fuller Test, Industrial Internet of Things (IIOT), Prediction, Forecasting, MAPE**

1. INTRODUCTION

Oil and Gas industry is one of the prominent and necessary industries. The workflow of these industries is generally categorized in three levels: (a) Upstream, (b) Midstream and (c) Downstream. Upstream part consists of the crude oil production and processing on the site itself and collecting the necessary data for analysis. Midstream part acts as an interface the production site and supplier chain managers. Downstream part is when the processed crude oil and gas reach the end shops and providers for consuming purpose.

There has been gradual rise in demand for IIOT implementation and analytics in Oil and Gas industries. Companies that invest funds to these kinds of industries want a reliable method to ensure progress and upliftment in production stats. Even a small amount of downtime of these oil and gas wells may result in big amount of loss both in terms of oil and gas production and in terms of monetary value industry. Faults may be of various kinds such as leakage of oil or gas, tampering of pipeline, sudden rise in temperature or pressure, pipelines valves malfunctioning causing sudden irregularities in sensor readings. So, companies need to be utmost sure that the wells function properly and if by any chance, faults tend to happen to any machinery equipment like in the ESP (Electronic Submersible Pump) or in the pipelines, we want that to be known beforehand. This is made possible using forecasting mechanisms such that if the actual value does not lie within our defined threshold boundaries (that we predicted) then we can pretend that the actual oil production is not up to the mark and there might be some wear and tear somewhere among the equipments that may be hampering our oil production. For the gathered time series, we need analysis to be done to forecast the oil gas or liquid generation and study the forecasted visually.

In upcoming sections, we will have Section 2 Literature survey and our problem statement, Section 3 Conceptual overview of algorithm used, Section 4 Implementation and Results. Section 5 Conclusion.

2. LITERATURE SURVEY

Analytics on oil and gas production data is done using variety of algorithm and regression techniques. Exponential smoothing models has been implemented [1] and XGBoost, also known as Extreme Gradient boosting is used [2] to forecast future predictions. Each algorithm is unique in its own domain and one need to analyse the data that is being worked upon before actually finalizing what method to choose. Sometimes Best algorithms tend to perform very poorly on certain datasets due to failure of recognizing existing patterns.

Various visual features that can be seen in time series plotting are:

(1) Trend – Trend can be positive or negative. Positive trend implies the graph has a positive slope over time and negative trend implies the plot goes downwards over time.
(2) Seasonality – Data is said to be seasonal if it exhibits a periodic trend like daily, monthly, quarterly or

yearly. Plot will show rise or fall upon reaching each period.

(3) Cycle – Data is said to have cyclic nature if its rises and falls are not of fixed and constant period.

3. ARIMA Model

ARIMA is a general class of forecasting model which have random trend, random walk, exponential smoothing, and autoregressive models as special cases. It is suitable for application on time-series data which can be made stationary by applying differencing or simple non-linear transformations like logarithm, deflation etc.

A time series dataset is stationary if its statistical properties are all constant over time. A stationary series has no trend, its variations around its mean have a constant amplitude, and it wiggles in a consistent fashion, i.e., its short-term random time patterns always look the same in a statistical sense.

ARIMA stands for Auto Regressive Integrated Moving Average. Here Autoregressive terms (AR) depend on the lags of the stationarized data, Moving Average (MA) depends on lags of the forecasted errors and a time series which needs to be differenced to be made stationary is said to be an "integrated" version of a stationary series.

A nonseasonal ARIMA model is denoted by ***ARIMA (p, d, q)***, where:

'p' is the number of autoregressive terms, 'd' is the number of non-seasonal differences suitable for stationarity, 'q' is the number of lagged errors in the prediction equation.

The forecasting equation is constructed as follows. First, let y denote the d^{th} difference of Y, which means:

If d=0: $y_t = Y_t$

If d=1: $y_t = Y_t - Y_{t-1}$

In terms of y, the general forecasting equation is:

$$\hat{y}_t = \mu + \phi_1 y_{t-1} + \ldots + \phi_p y_{t-p} - \theta_1 e_{t-1} - \ldots - \theta_q e_{t-q} \qquad \ldots (i)$$

μ is the average of the series,
ϕ_1, ϕ_2 ... are AR parameters denoted as AR (1), AR (2) ...,
θ_1, θ_1... are MA parameters denoted as MA (1), MA (2) ...

4. Implementation and Results

The dataset we are using in this paper is the heat sensor data from an active oil and gas well.

Heat-tracing temperature sensors are made for use in systems that measure the surface temperature of process pipes that is carrying products whose temperatures must be controlled to prevent freeze-up, or to maintain a viscosity level so that the inner medium will flow.

We have taken a small subset of time series data, recorded every 10 minutes over a span of 9 days.

Let us plot the our original dataset and analyse the graphs. (Fig 1). x-axis denotes the time from 1 Jan 2018 to 11 Jan 2018 and y-axis denotes the temperature.

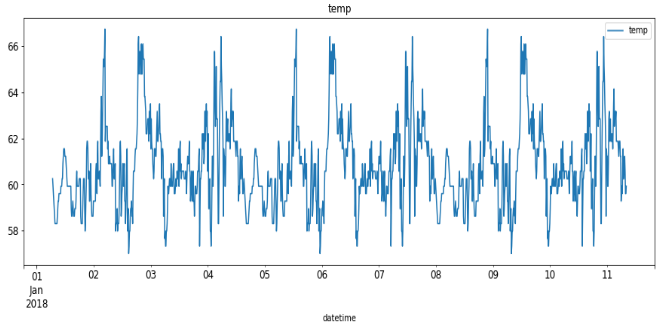


Fig 1. Time series plot of the dataset

This plot looks complex for comprehension. Let us resample it hourly to get an appropriate plot for study. Fig 2 shown the resampled dataset.

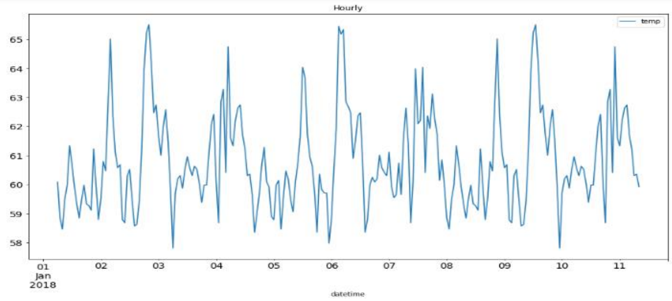
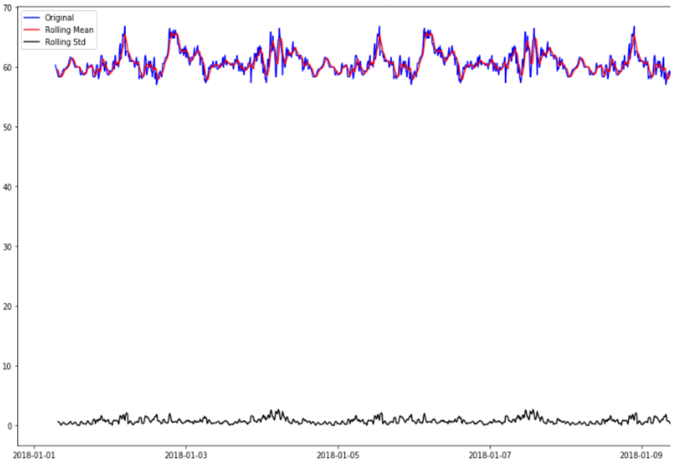


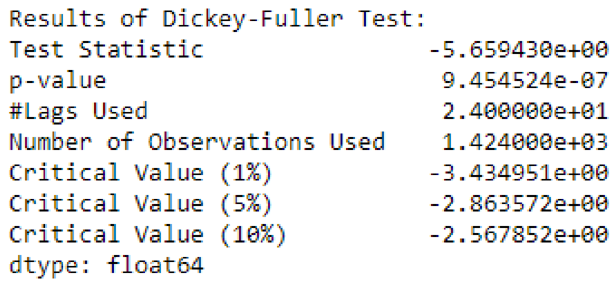
Fig 2. Resampled dataset hourly

Now we check whether series is stationary or not. We always pre-process the dataset to attain as much stationarity as possible. There are two ways to check stationarity:

(A) **Plotting Rolling statistics**: In this test, we plot the moving average and analyse how much variance is there in the plot.
(B) **Dickey-Fuller Test**: This is one of the statistical tests for checking stationarity. Here the null hypothesis is that the TS is non-stationary. The test results comprise of a ***Test Statistic*** and some ***Critical values*** for difference confidence levels. If the 'Test Statistic' is less than the 'Critical Value', we can reject the null hypothesis and say that the series is stationary.
Now let's plot the Rolling Average and visually analyse whether data is stationary. Also applying dickey fuller for hypothesis testing.

Now let's plot the Rolling Average and visually analyse whether data is stationary. Also applying dickey fuller for hypothesis testing.





```
Results of Dickey-Fuller Test:
Test Statistic                 -5.659430e+00
p-value                         9.454524e-07
#Lags Used                      2.400000e+01
Number of Observations Used     1.424000e+03
Critical Value (1%)            -3.434951e+00
Critical Value (5%)            -2.863572e+00
Critical Value (10%)           -2.567852e+00
dtype: float64
```

Fig 3. Moving Average statistics and Dickey Fuller test for testing stationarity

Trend seems to be stationary by seeing the test statistic because test statistic is much lower than the critical values. However, there may be seasonality component that we might erroneously miss.

One of the methods to remove trend and seasonality is to take logarithm of the series. As a result, Trend, if present is very much smoothened out and the seasonality gets reduced as well. Below is the moving average plot for the log valued dataset shown in Fig 4.

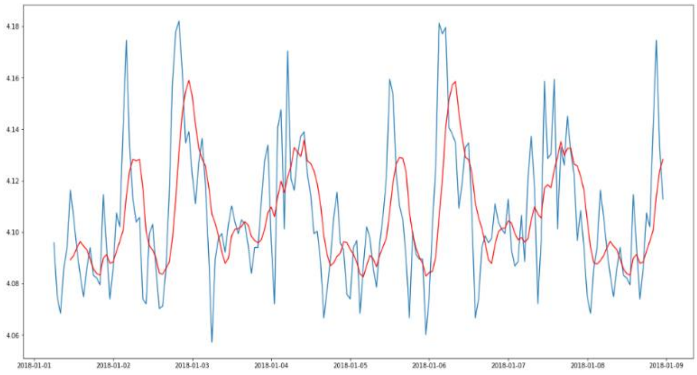
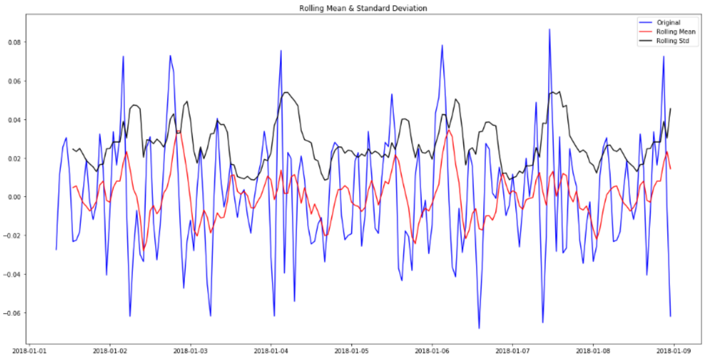
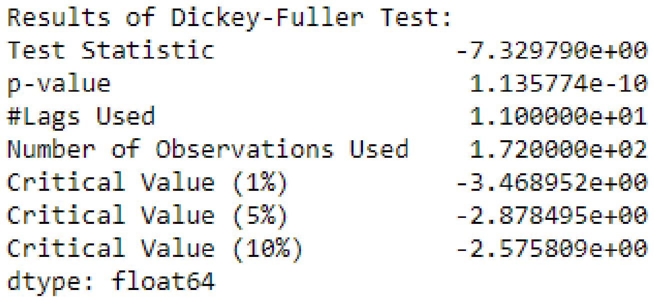


Fig 4. Moving average plot of logged series

Now we suspect that there exists some seasonal characteristics. To check that, we perform first order differencing. Fig 5 shows the differenced plot of the 'logged' series.





```
Results of Dickey-Fuller Test:
Test Statistic                 -7.329790e+00
p-value                         1.135774e-10
#Lags Used                      1.100000e+01
Number of Observations Used     1.720000e+02
Critical Value (1%)            -3.468952e+00
Critical Value (5%)            -2.878495e+00
Critical Value (10%)           -2.575809e+00
dtype: float64
```

Fig 5. Differenced 'logged' series statistics

From *Fig 5*, we see that the p-value reduced drastically. Thus, we can say that we have made our dataset more stationary by first performing log of the series and the differencing by lag 1 **(d = 1)**.

It is time to predict values of AR order (p) and MA order(q). We use two plots to determine these numbers. Let's discuss them first.

Autocorrelation Function (ACF): It is a measure of the correlation between the Time series with a lagged version of itself. For instance, at lag 5, ACF would compare series at time instant 't1'…'t2' with series at instant 't1-5'…'t2-5' (t1-5 and t2 being end points).

Partial Autocorrelation Function (PACF): This measures the correlation between the TS with a lagged version of itself but after eliminating the variations already explained by the intervening comparisons. Both the ACF and PACF plots are shown in Fig 6.

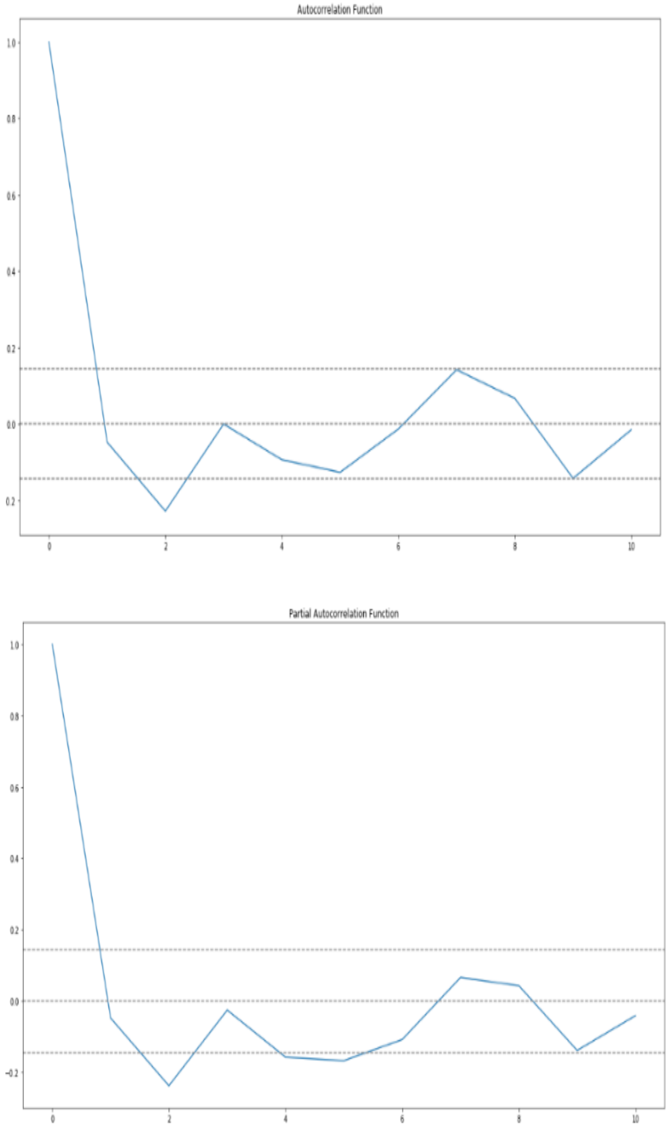


Fig 6. ACF and PACF plots for prediction of 'p' and 'q' order

The dotted line up and down of the zero y-value are the confidence intervals.

'p' is determined by checking where the PACF curve first crosses the upper confidence level.

'q' is determined by checking where the ACF curve first crosses the upper confidence level.

So, by looking at Fig 6 we conclude that **(p = 1)** and **(q = 1)**.

Now, we fit ARIMA using combination of these two parameters see the model with the least RSS (Residual sum of squares).

Red curve is the fitted curve, **Blue** curve is the original data

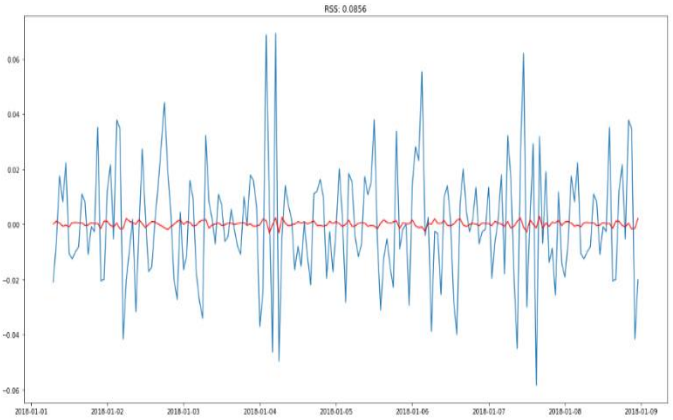


Fig 7. ARIMA (1,1,0) fitting with RSS=0.0856

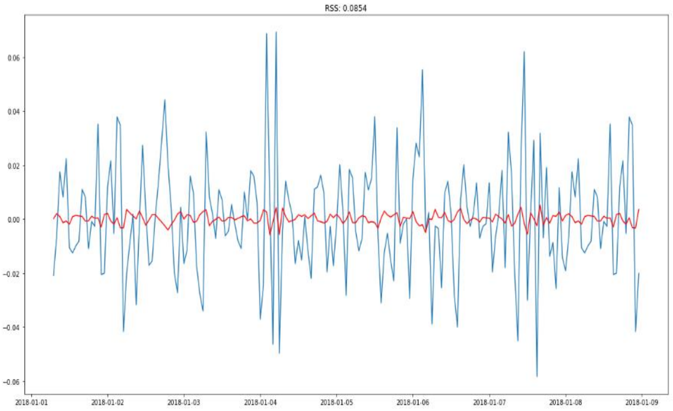


Fig 8. ARIMA(1,1,0) fitting with RSS=0.0854

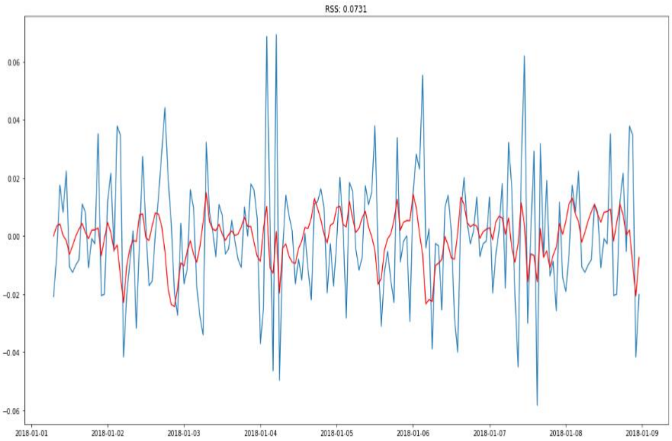


Fig 9. ARIMA(1,1,1) fitting with RSS=0.0731

By analysing Fig {7}, {8}, {9}, we see that ARIMA (1,1,1) has the least RSS value of 0.0731. So, this ARIMA with parameters p = 1 & q = 1 is more suitable compared to others. Finally, we used ARIMA (1,1,1) for forecasting the data. We forecasted up to 4 days in future i.e. from 2018-01-09 to 2018-01-13.

The red dotted line shows the boundary limit of confidence interval on '2018-01-11'.

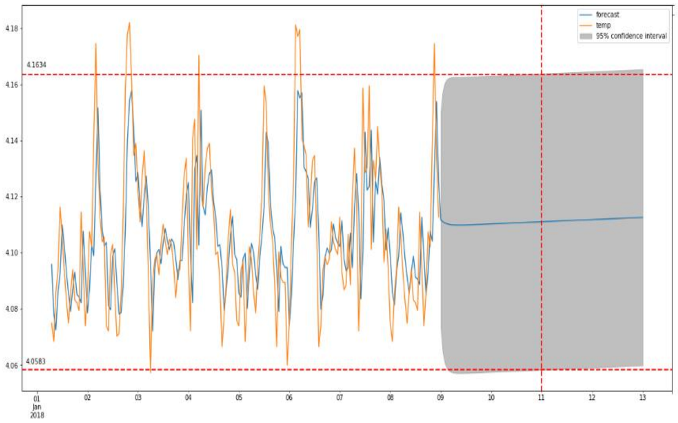


Fig 10. Forecasted confidence band till '2018-01-13'

Upper limit value is **4.1634** and lower limit is **4.0583.** Now since we applied ARIMA on logged valued dataset (log is taken to remove any existing trend and seasonality). So, we must convert the values back into original form. For that we perform exponentiation of these upper and lower limits.

$e^{4.1634}$ **= 64.28974** (upper limit), $e^{4.0583}$**= 57.87584** (lower limit)

So, the predicted value on '2018-01-11' should lie within this limit. In both cases, if measured temperature goes beyond upper limit or drops below lower limit, then the appropriate alert team or onsite supervisor will be alerted that there might happen some malfunctioning in equipments due to which the values are shown to be abnormal.

5. CONCLUSIONS

The risk assessment of IIOT equipments are of utmost importance to industries because malfunction for even a fraction of time in the Oil and gas drilling equipments such as ESP (electronic submersible pumps) and pipelines may lead to heavy loss both from production as well as economic point of view.

This paper showed a decent approach to prevent such scenario by analysis the data sent by sensor like temperature, well head pressure etc and watch out for anomalies in real time by forecasting using ARIMA generic model. Now if at any time in the forecasted future, we see an anomaly among predicted dat values, we can alert the site engineer prior to any malfunction that could occur in the equipments so that corrective measures may be taken to prevent possible risk associated with the onsite machinery.

REFERENCES

[1] Wen ZHANG, Zhansheng SONG, Qingping Wang, Weidong HAO The Forecasting of Workload of Oil Production Program,2011 International conference of business management, pp.452-456.

[2] Mesut Gumus, Mustafa S. Kiran, Crude oil price forecasting using XGBoost,2017 conference of computer science and engineering, pp.1100-1102

[3] Wazir Khan, Muhammad Aslem, Khurram Khan, Shoaib Hussain, 2017. A reliable
Internet of Things based architecture for oil and gas industry: International
Conference on advanced computing technology. DOI: 10.23919/ICACT.2017.7890184

[4] Shamisa Shoja, Aliakbar Jalali, "A study of the Internet of Things in the oil and gas industry, 2017. Proceedings., IEEE International Conference on knowledge based
engineering and innovation.16(1), pp.414-454

[5] Pan Yi, Lizhi Xiao, Yuanzhong Zhang "Remote real-time monitoring system for oil and gas well based on wireless sensor networks, June 2010.17. pp. 55-83. ISSN-11461861

[6] Fred Florence, December 2013. Upstream oil and gas drilling processes and instrumentation opens to new technology. IEEE Instrumentation and

measurement magazine, Volume 16, Issue 6, ISBN: 978-89-968650-9-4

[7] When Zhang, Zhenshen Song and Quinping Wing, Juan Carlos and Clark, (2011), The forecasting of workload of oil production program. IEEE International Conference on
Business management and Electronic Information.15(8), ISBN: 978-1-61284-109-0

[8] Guojian Cheng, YaoAn, Zhe Wang, Kai Zhu, "Oil Well Placement Optimization using Niche Particle Swarm Optimization". 2012 IEEE International Conference of Advanced Engineering.

[9] Mariana Araujo, Jose Aguilar, Hugo Aponte, "Fault Detection System in Gas Lift Well based on Artificial Immune system", IEEE 2003,5, pp.114-127.

[10] Hanyeu Zhang, Cheng Fei. "Design of Oil Well Monitoring Information Management System Based on IOT Technology" Springer Publication, December 2015.

BIOGRAPHIES

Anurag Kumar Singh, currently pursuing MTech in Advanced Computing from Maulana Azad National Institute of technology, Bhopal. He did Bachelor of Engineering from Bhilai Institute of technology, Durg

Dr. R.K. Pateriya is an Associate professor in Maulana Azad National Institute of technology, Bhopal. He holds PHD(CSE), MTech(CSE) and BE(Computer technology) degrees and has guided many M.Tech and PhD scholars.