EECS E6690 – Final Project

Rohan Mehta (rm3500)

Nanda Kishore Siddabasappa (ns3212)

Statistical Analysis of Cloud Datacenter Workloads

Introduction

- “Forecast” is very “cloudy” for next few years!
- Need to understand the underlying workload characteristics in order to efficiently plan and manage the resources.
- The traces collected contain information about memory, CPU, network I/O, and disk I/O.
- Models and techniques used for forecasting:
  - Polynomial Regression
  - ARIMA
  - Prophet (by Facebook)

Data Set

- The dataset contains the performance metrics of 1250 VMs from a distributed datacenter from *Bitbrains* for one month duration at interval of 5 minutes for each observation (i.e. approximately 8600 observations per VM).
- The format of each file is row-based. Each row represents an observation of the following 11 performance metrics:

Features	Description
Timestamp	number of milliseconds since 1970-01-01
CPU cores	number of virtual CPU cores provisioned
CPU capacity provisioned	the capacity of the CPUs in terms of MHZ, it equals to number of cores * speed per core
CPU usage	in terms of MHZ
CPU usage	in terms of %
Memory provisioned	the capacity of the memory of the VM (in KB)
Memory usage	the memory that is actively used in terms of KB.
Disk read throughput	in terms of KB/s
Disk write throughput	in terms of KB/s
Network received throughput	in terms of KB/s
Network transmitted throughput	in terms of KB/s

Pedagogy

Reproduction of some results from reference paper

Data analysis and feature selection (5 out of 11)

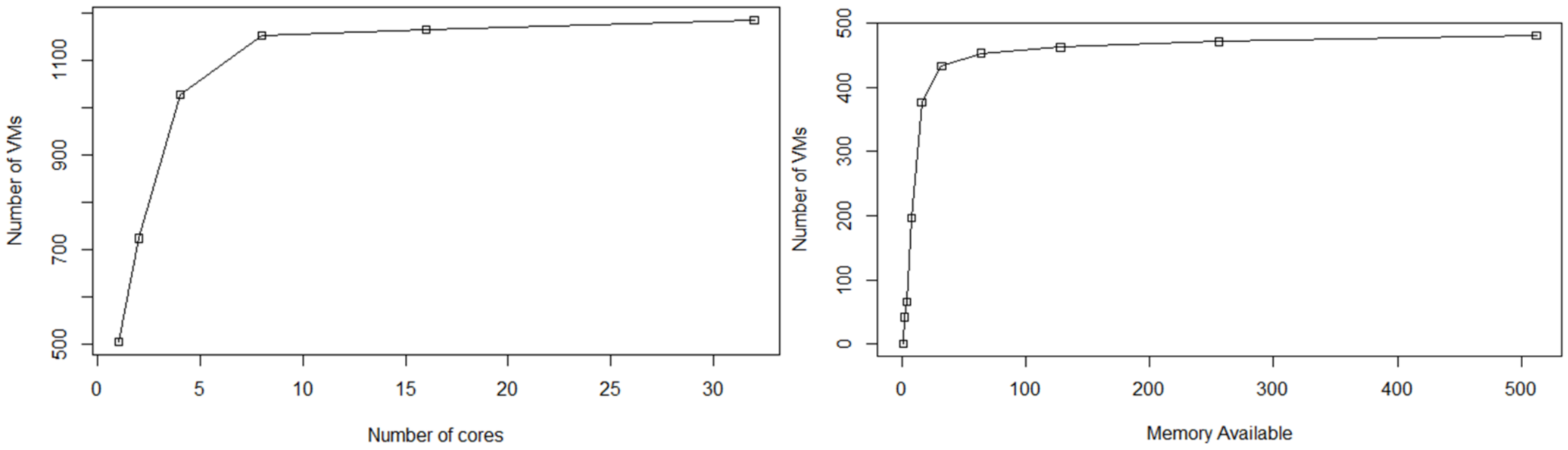
Cleaning of data

Applied 3 models to each of the four features:

* Polynomial Regression
* ARIMA
* Prophet (by Facebook)

Analysis of results obtained from these models

Reproduction of Results from Reference Paper

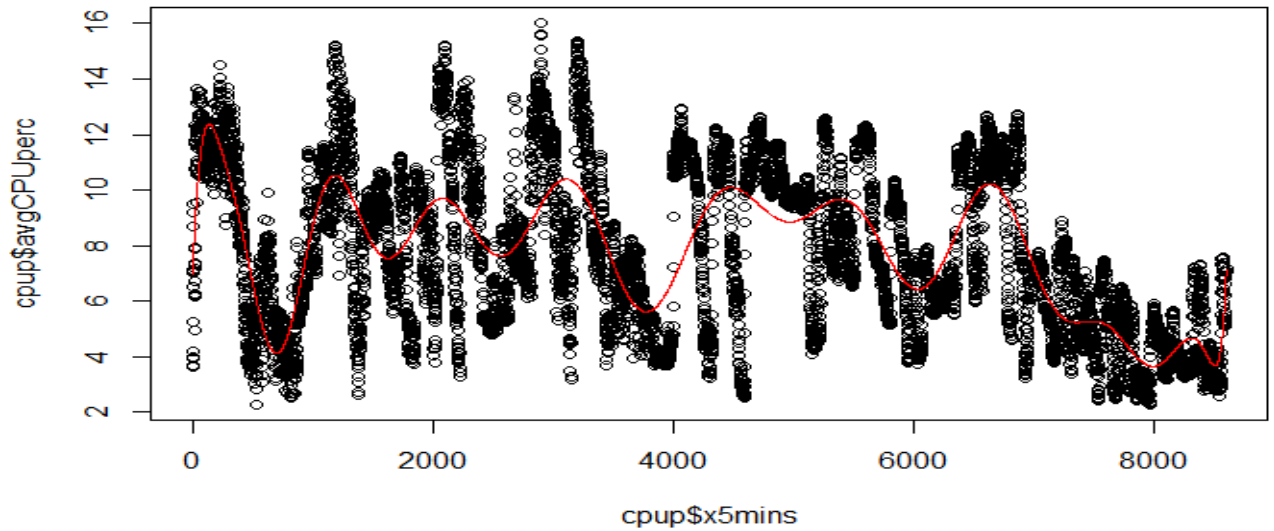


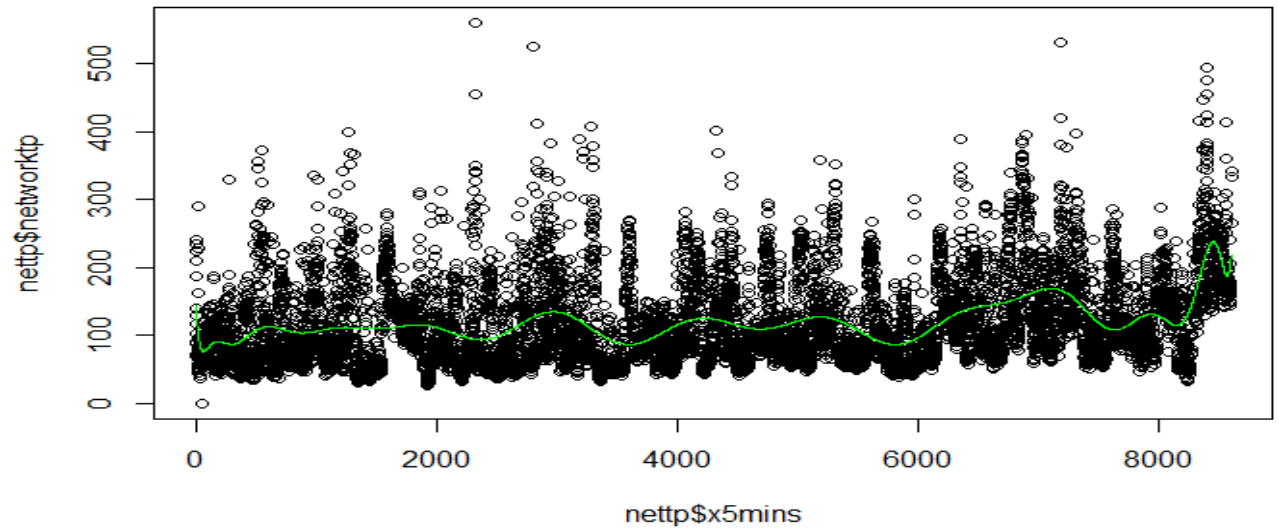
CDF of number of requested CPU cores (left) and amount of requested memory (right)

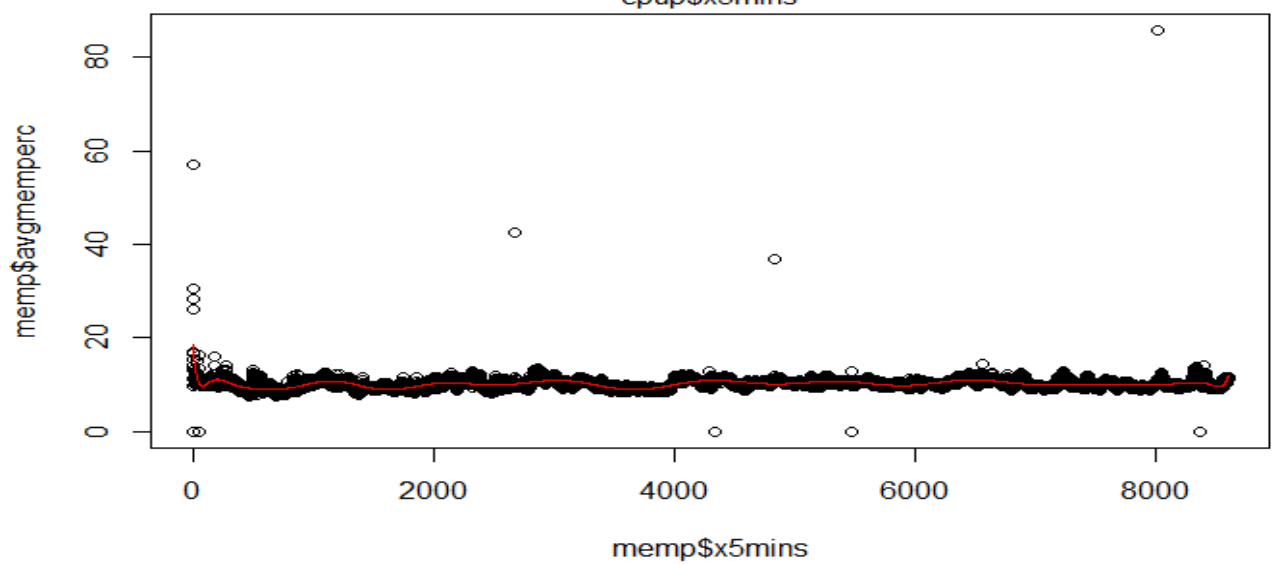
New Approach

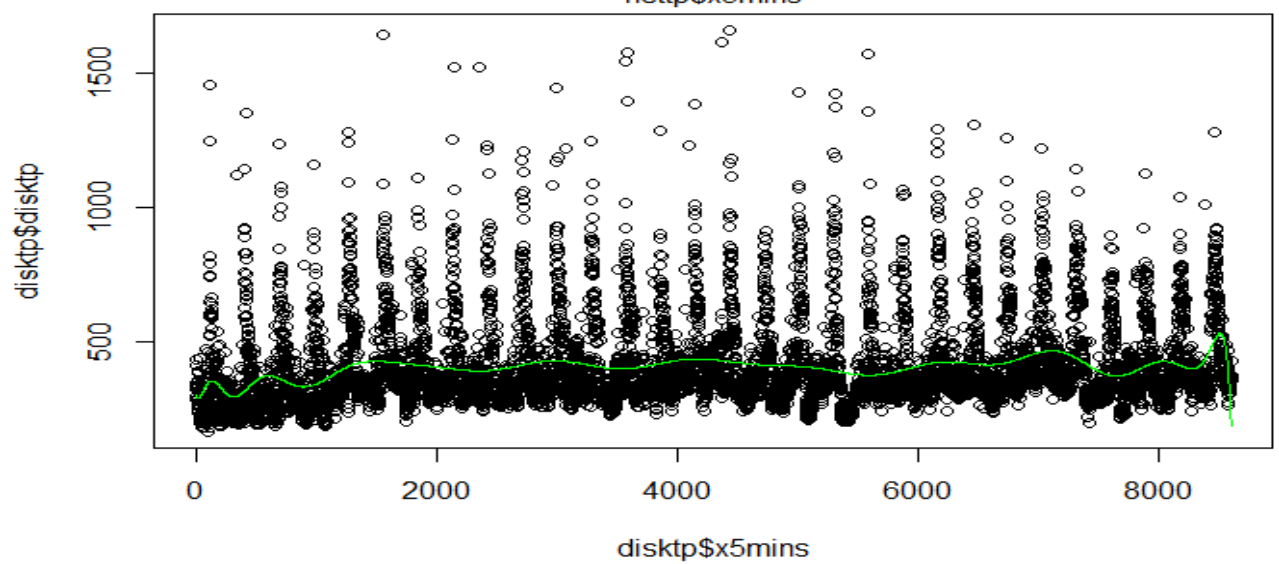
- We have focused our problem on studying following four features over one month period as a time series:
  - Percentage CPU usage
  - Percentage memory usage
  - Total Disk I/O
  - Total network bandwidth I/O
- We applied some statistical learning models on each of these four parameters over a period of time and then figured out which model works best for a particular parameter.
  - Polynomial Regression
  - ARIMA
  - Prophet (by Facebook)

Polynomial Regression

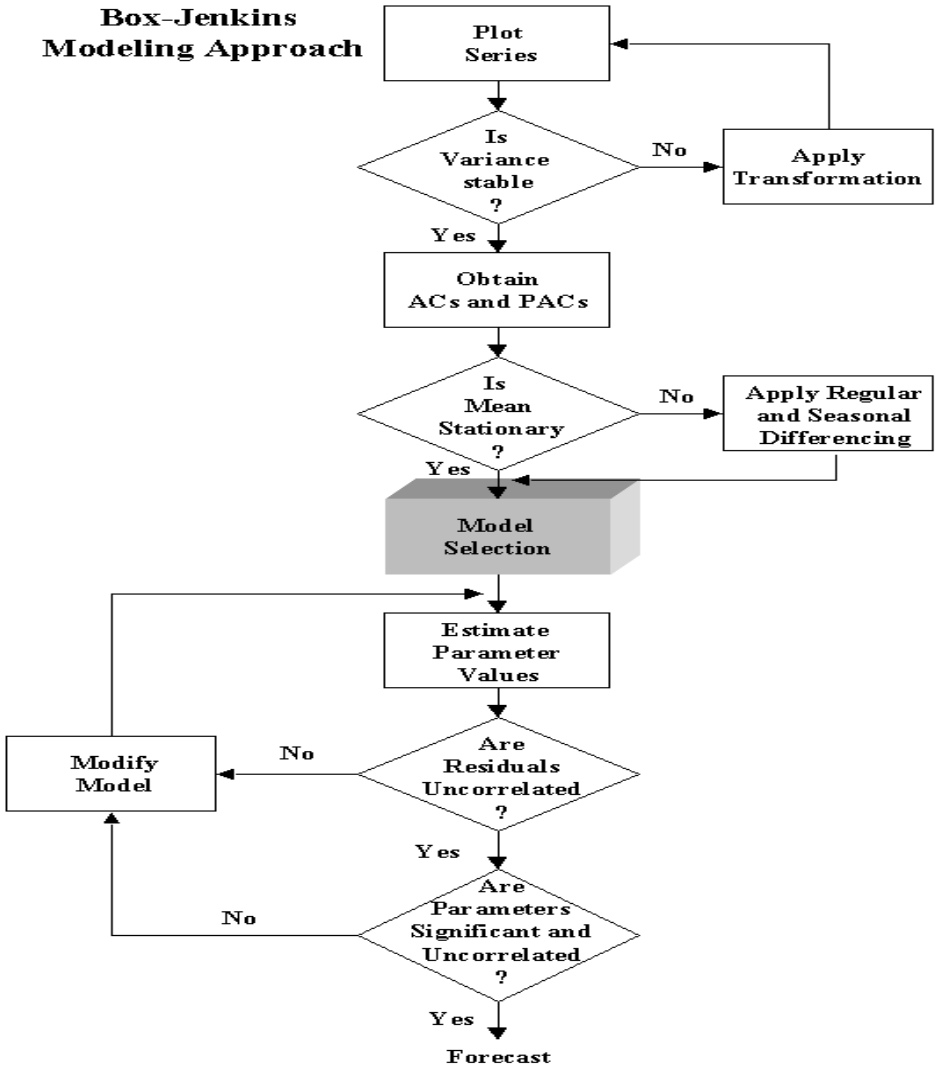


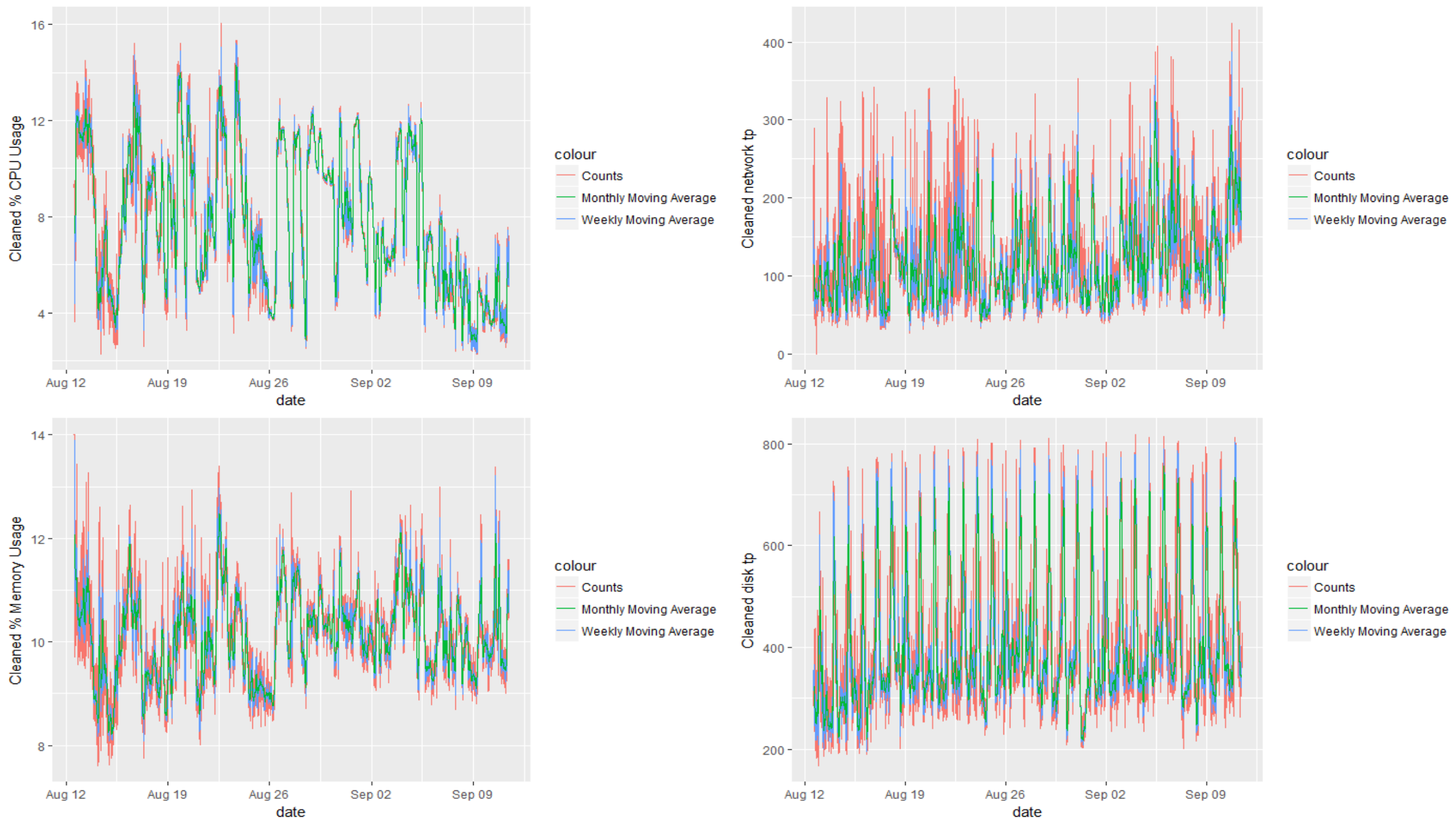


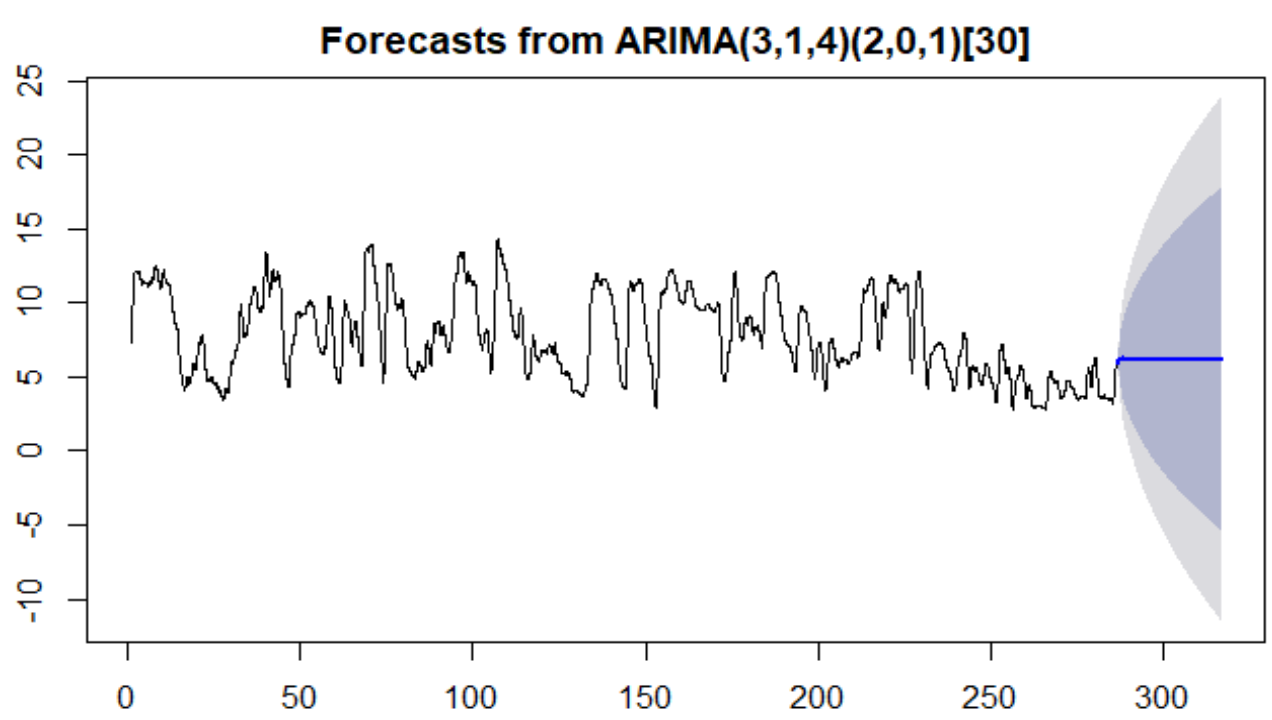


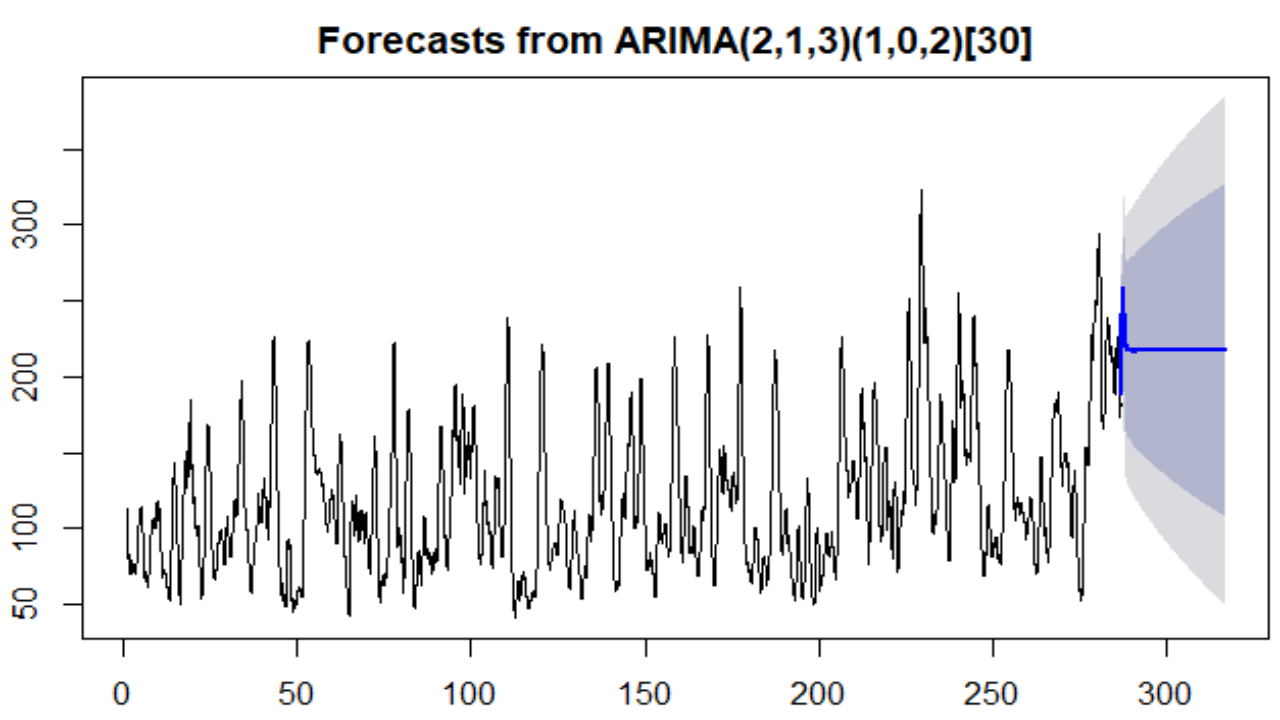


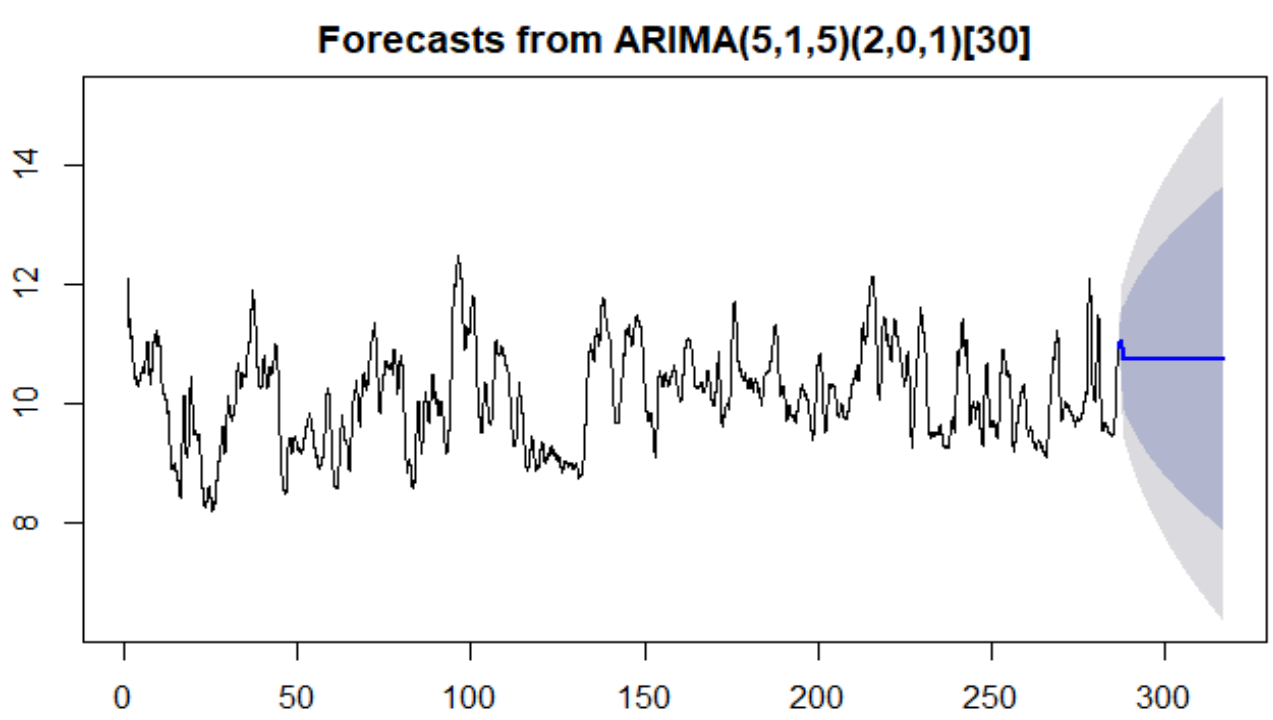
Autoregressive Integrated Moving Average (ARIMA)

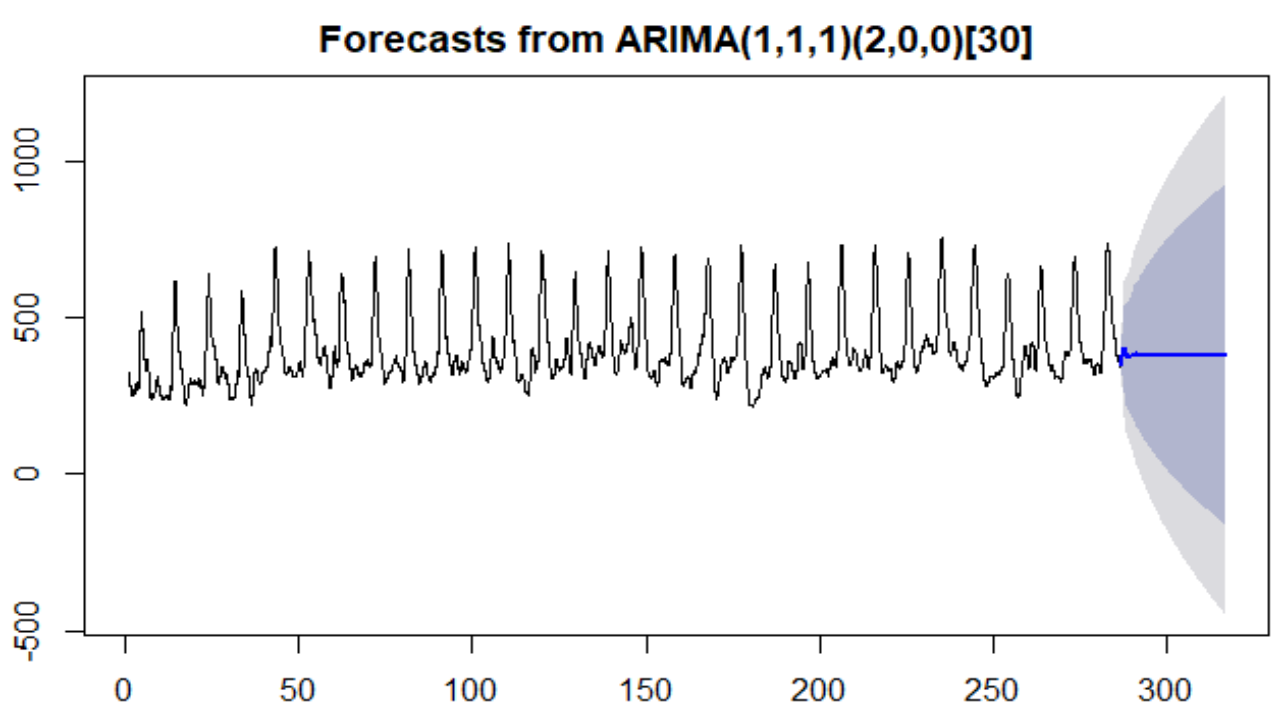



Cleaned % CPU Usage
16
12
8
4
Aug 12
Aug 19
Aug 26
Sep 02
Sep 09
date
colour
Counts
Monthly Moving Average
Weekly Moving Average
Cleaned network tp
400
300
200
100
0
Aug 12
Aug 19
Aug 26
Sep 02
Sep 09
date
colour
Counts
Monthly Moving Average
Weekly Moving Average
Cleaned % Memory Usage
14
12
10
8
Aug 12
Aug 19
Aug 26
Sep 02
Sep 09
date
colour
Counts
Monthly Moving Average
Weekly Moving Average
Cleaned disk tp
800
600
400
200
Aug 12
Aug 19
Aug 26
Sep 02
Sep 09
date
colour
Counts
Monthly Moving Average
Weekly Moving Average

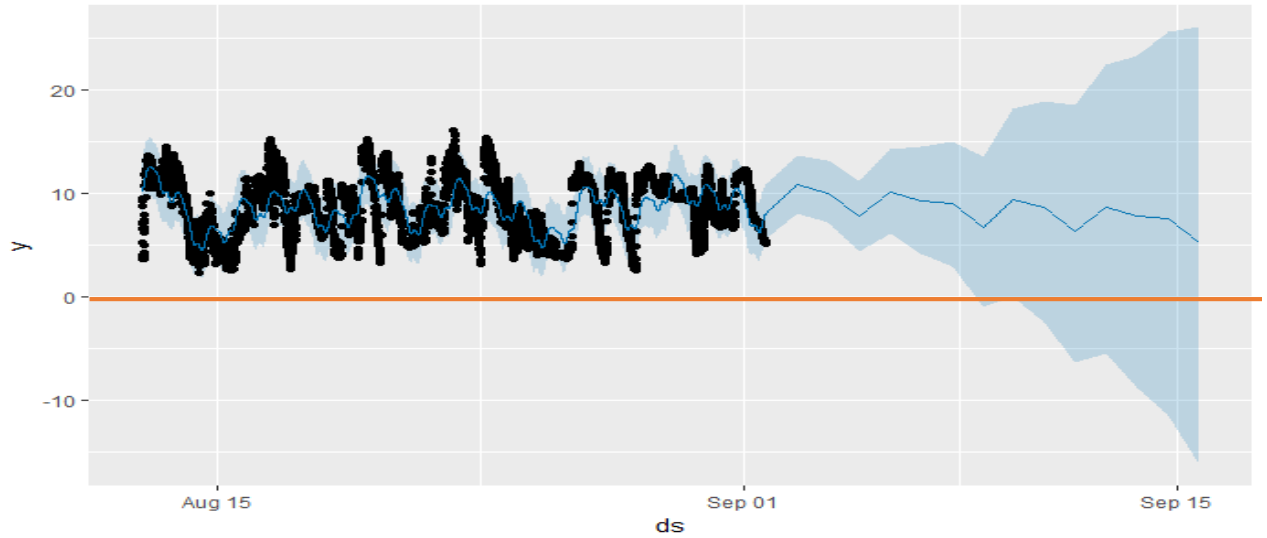

Forecasts from ARIMA(3,1,4)(2,0,1)[30]
25
20
15
10
5
0
-5
-10
0
50
100
150
200
250
300

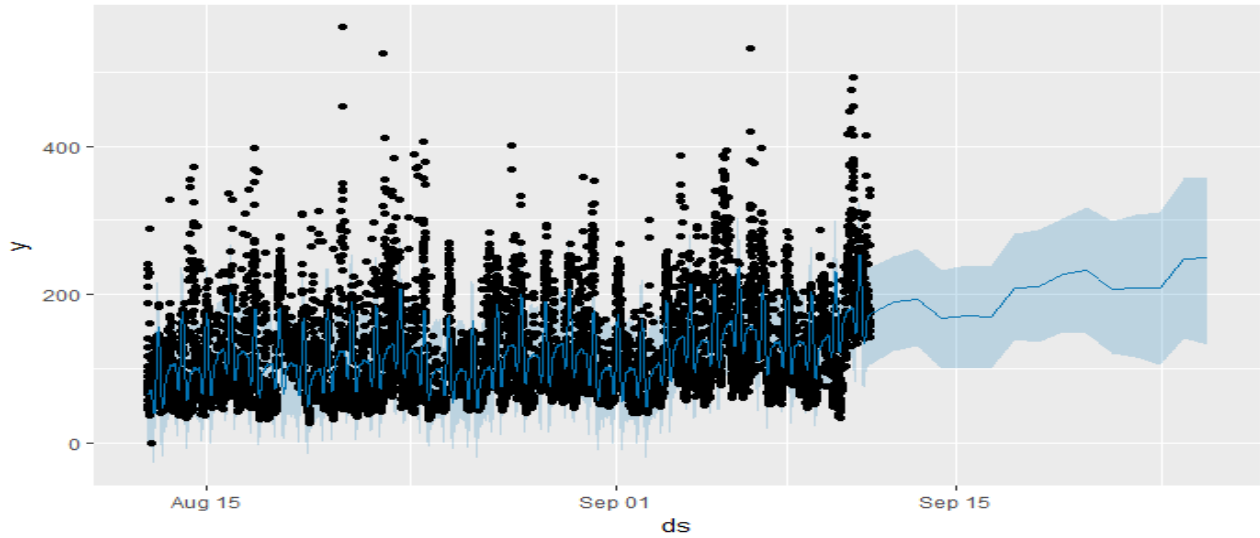

Forecasts from ARIMA(2,1,3)(1,0,2)[30]
300
200
100
50
0
50
100
150
200
250
300

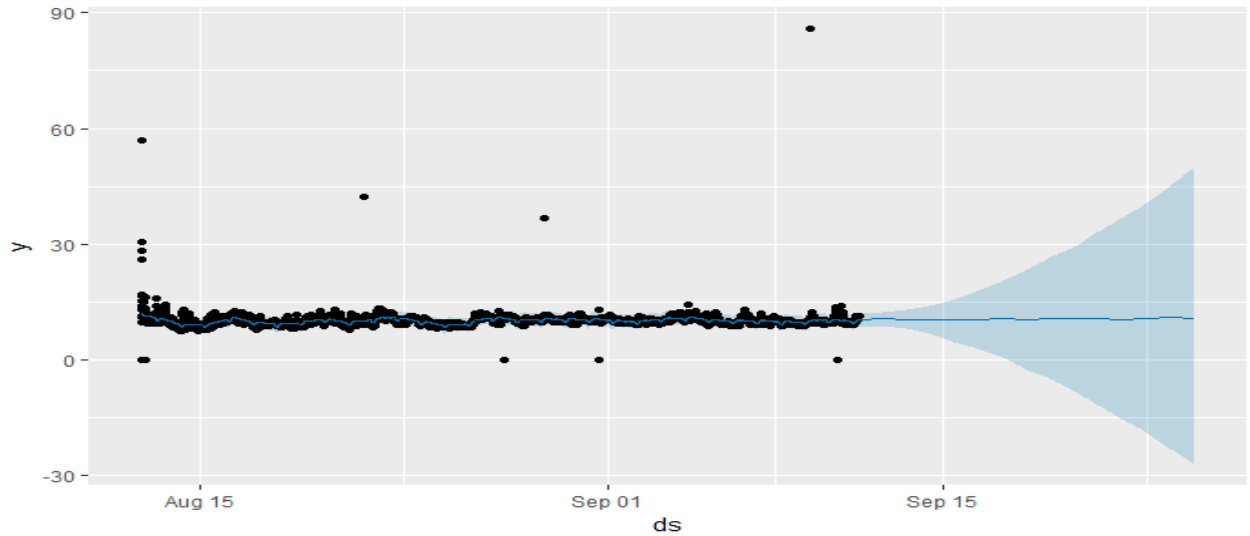

Forecasts from ARIMA(5,1,5)(2,0,1)[30]
14
12
10
8
0
50
100
150
200
250
300

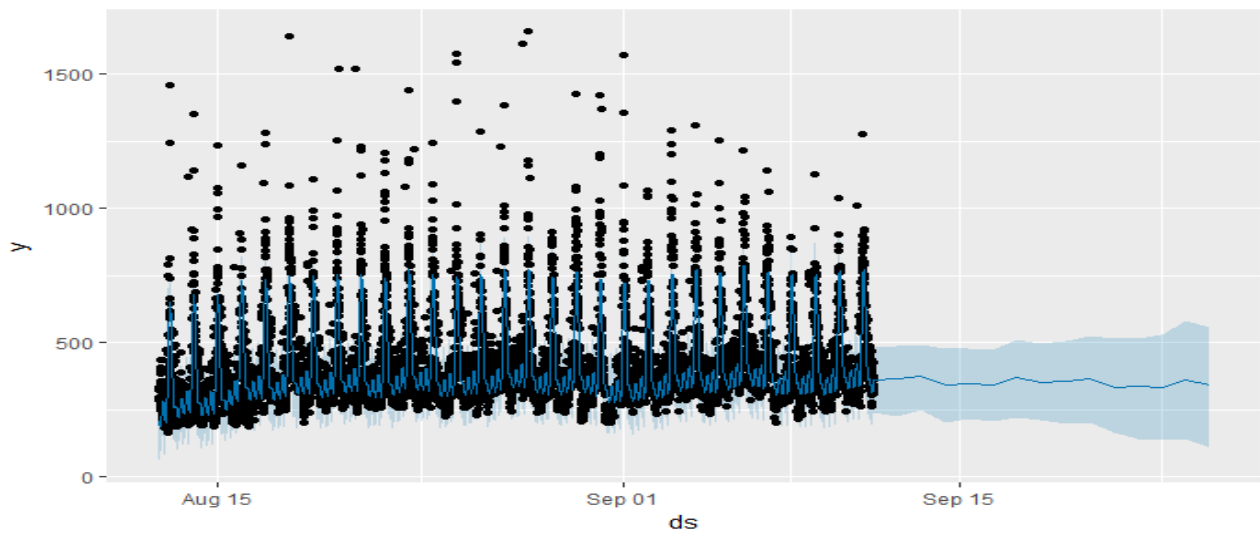

Forecasts from ARIMA(1,1,1)(2,0,0)[30]
1000
500
0
-500
0
50
100
150
200
250
300

Prophet


20
10
0
-10
y
Aug 15
Sep 01
Sep 15
ds


400
200
0
y
Aug 15
Sep 01
Sep 15
ds


90
60
30
0
-30
y
Aug 15
Sep 01
Sep 15
ds


1500
1000
500
0
y
Aug 15
Sep 01
Sep 15
ds

Conclusion

- Polynomial regression fits nicely as we increase the order. However, this gives us absurd and impracticable future predictions. Clearly, this can be explained by the phenomenon of *overfitting*.
- Prophet model generally outperforms ARIMA model when it comes to predicting future resource usage.
  - It is also less complicated to use as compared to ARIMA.
- We need more data, preferably at least for 1 year, so that we can make more accurate predictions for resource utilization.
  - Prophet model usually advocates to use data of 1 year duration.

Future Work

- If more data is available, we can do similar analysis for different topological models and get insights on resource management principles.
- We can try different time-series forecasting models (like Neural Networks, Exponential Smoothing, etc.) as alternative learning techniques.
- Outliers/Irregularities in data can be studied further to identify types of malicious attacks in a cloud datacenter (e.g. DDoS).

References

1. Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. An Introduction to Statistical Learning with Applications in R. 2nd edition, Springer, 2009.
2. Siqi Shen, Vincent van Beek, and Alexandru Iosup. *Statistical Characterization of Business-Critical Workloads Hosted in Cloud Datacenters*, the 15th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing (CCGrid), 2015, Shenzhen, China
3. Lecture slides
4. https://www.rdocumentation.org/
5. http://gwa.ewi.tudelft.nl/datasets/gwa-t-12-bitbrains - data set
6. https://www.datascience.com/blog/introduction-to-forecasting-with-arima-in-r-learn-data-science-tutorials - ARIMA model
7. https://research.fb.com/prophet-forecasting-at-scale/ - Prophet Model
8. https://towardsdatascience.com/using-open-source-prophet-package-to-make-future-predictions-in-r-ece585b73687 - Prophet Model
9. Sean Taylor and Benjamin Latham. *Forecasting at Scale*, PeerJ Preprints, https://doi.org/10.7287/peerj.preprints.3190v2 , CC BY 4.0 Open Access, 27 Sep 2017

Thank you!

Questions?