SINTECHCOM: Science, Technology and Communication Journal
Vol. 2, No. 2, February 2022, pp. 35-48, DOI: 10.59190/stc.v2i2.207

Integrated production facilities clustering and time-series forecasting derived from large dataset of multiple hydrocarbon flow measurement

Adityapati Rangga[1],*, Yohana Dewi Lulu Widyasari[2], Dadang Syarif Sihabudin Sahid[2]
[1]Department of Computer, Politeknik Caltex Riau, Pekanbaru 28265, Indonesia
[2]Department of Information Technology, Politeknik Caltex Riau, Pekanbaru 28265, Indonesia

ABSTRACT

In the complex, mature, and large oilfields, there is a need for Integrated solution in order to have a helicopter view of entire facilities throughput. The real time metering information provides an on-demand daily data and trend. However, it is rarely being connected to analytics solution for business intelligence such as, prediction, optimization, decision support and forecast. This paper cover about exploratory data analysis of large dataset of multiple hydrocarbon facilities metering within integrated network, performing multi-feature data clustering and making a time-series forecasting techniques. K-means and PCA are combined to make cluster of production facilities which resulted with gas processing cluster, high oil producer, high water processing station, and the lowest performer in term in hydrocarbon processing. Furthermore, VAR and LSTM are compared as forecasting tools for day-to-day fluid prediction, to maintain normal operational scenario.

ARTICLE INFO

Article history:

Received Dec 26, 2021
Revised Jan 28, 2022
Accepted Feb 10, 2022

Keywords:

Clustering
Forecast
LSTM
Oilfields
VAR

* **Corresponding Author**
E-mail address: adityapati20s2tk@mahasiswa.pcr.ac.id

1. INTRODUCTION

Production networks and processing facilities can be very complex, with multiple interactions and constraints. Start from the reservoir fluids whose properties, such as gas-oil ratio, gas density, water cut, are changing with time, flowlines, gas and liquid handlings capacity and constraints [1-3]. It is required to have representative number of current performances, predict the state, and forecast the fluid according to the historical performance [4, 5].

There have been many documented applications of such production optimization and modelling, ranging from reservoir management, through well work, offline data-driven and physics-based modelling to advanced control [6-8].

Mixed integer nonlinear programming (MINLP) seems able to answer, the optimal routing, accommodate the detail operating mode in physical system and plant component, for optimum configuration, however it is not based on historical data performance [9-11]. It is most likely focus on the detail of the component of the systems, not the entire operation and production performance in a unique multiple and scattered gathering station facilities [12, 13]. It is frequently well test-based per field, and not measurement based per gathering station.

The oil and gas standards community has been working to enable end-to-end, real-time data transfer from the sensor to simulation or the accounting system or a regulator, in addition to the traditional exchange of larger static information [14, 15]. The industry has seen a great deal of progress on the uptake of data standards for data in motion and fundamental improvements in the standards themselves. The data in motion are real-time data in the drilling and production arenas, and on-demand movement of data between applications or among partners and regulators [16-18]. The static data is the traditional contextual information about wells and their histories along with information on their historical performance and the activities used to create and operate them [19, 20].

Integrated combination of production clustering, throughput and forecasting are necessary in a geographically scattered facility [21, 22]. Resources need to be maintained, developed and prioritized according to the priority cluster, and the production fluctuation need to be addressed in advance, to anticipate the short-term fluid fluctuation, and as an addition to surveillance and optimization activity.

Hence, this research will propose the utilization of big-historical-data performance of hydrocarbon fluid output, from multiple gathering station, answering the current state performance classification through K-means, forecasting the future with VAR and LSTM, and leading to accurate prioritization and decision support in competitive business environment point of view.

2. MATERIALS AND METHOD

The scope of such production optimization has usually been limited to the production network only, i.e., from the sand face to the separators: neither the reservoir nor the facilities have been explicitly modelled. Instead, these have been approximated using well performance curves, specified fluid properties and suitable constraints [1]. Although traditional model-based production optimization has already demonstrated significant benefits, the use of fixed constraints to represent the reservoir and/or facilities is prone to error.

In order to overcome these challenges, Woodman et al. (2017), use MINLP optimization [1]. Hence, it is still provided complexity by including a lot of components of production facilities network, and try to reconcile the pressure-driven approach that used by production network model, with flow-driven approach that used by process simulator.

2.1. Hydrocarbon Facility

2.1.1. Gathering Station

Large oilfield typically consists of multiple gathering station facilities. A gathering station is a facility for processing the production fluid from oil wells to separate gas, water and oil. The main end product of the gathering station is crude oil that meet the basic sediment and water (BSW) requirement. Typical main processing unit on the gathering station is following:

- Gas separation unit.
- Water-oil separation unit.
- Oil handling facilities.
- Water handling facilities.
- Waste gas handling facilities.

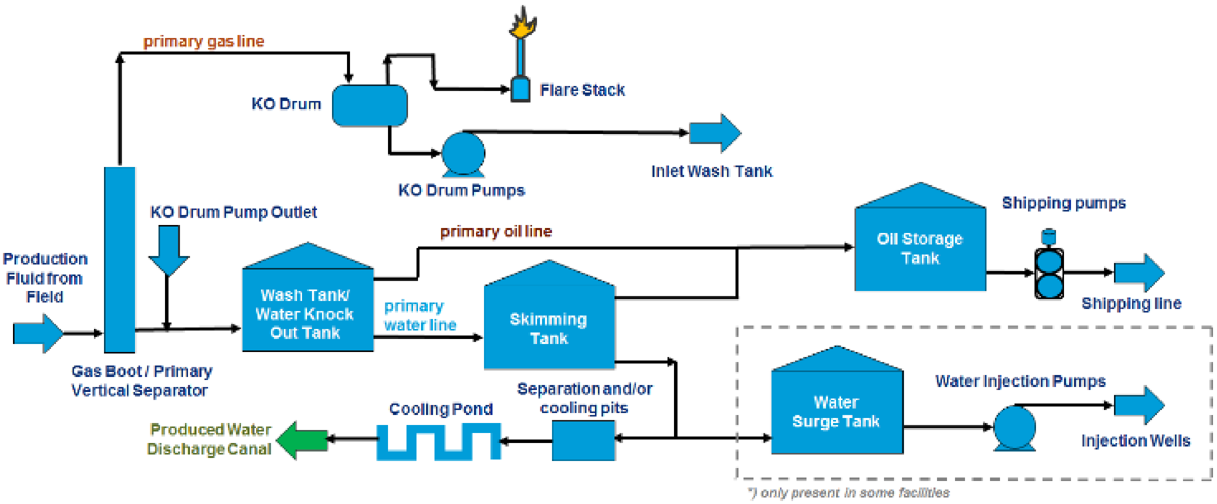


Figure 1. Typical gathering station facility.

The simplify of the schematic diagram for the process in the Gathering station is outlined in Figure 1.

2.1.2. Gas Plant

Gas Plant is processing raw gas coming from mostly gas wells and also from other source, that is an associated gas (gas separated from oil) from gathering station. The main process on the gas plant is liquid separation process, compression process and dehydration process. The final product from gas plant is a dry gas that need to meet internal requirement, delivered to gas turbine stations as fuel for generating electric power.

2.2. Hydrocarbon Production

According to facilities type, we can easily guess, there are 3 type of composite substance that extracted from the subsurface, processed in facility, and measured by metering systems. There are oil, gas, and water. In details, it can be listed as following,

- Oil, as a primary product of the oilfields.
- Produced water.
- Associated gas, natural gas produced by oil wells.
- Non-associated gas, natural gas produced by natural gas wells.
- Condensate, a natural gas liquid with a low vapor pressure compared with natural gasoline and liquefied petroleum gas [4].

2.3. Methodology

2.3.1. K-Means Clustering

The K-means algorithm (KM) partitions data into k sets. The solution is then a set of k centers, each of which is located at the centroid of the data for which it is the closest centre. For the membership function, each data point belongs to its nearest centre, forming a Voronoi partition of the data [6]. The objective function that the KM algorithm optimizes is:

$$KM(X,C) = \sum_{i=1}^{n} \underset{j\in\{1\ldots k\}}{min} \left\| x_i - c_j \right\|^2 \tag{1}$$

This objective function gives an algorithm which minimizes the within-cluster variance (the squared distance between each center and its assigned data points).

The membership and weight functions for KM are:

$$m_{KM}(C_l|X_i) = \begin{cases} 1\,; if\ l = \arg min_j \left\| x_i - c_j \right\|^2 \\ \quad 0\,; otherwise \end{cases} \tag{2}$$

$$w_{KM}(x_i) = 1 \tag{3}$$

KM has a hard membership function, and a constant weight function that gives all data points equal importance. KM is easy to understand and implement, making it a popular algorithm for clustering. The objective using K-Means is to perform unsupervised learning to production data, and revealed the most objective clustering to segregate the performance of production facilities.

2.3.2. Principal Component Analysis (PCA)

PCA is defined as an orthogonal linear transformation that transforms the data to a new coordinate system such that the greatest variance by some scalar projection of the data comes to lie on the first coordinate (called the first principal component), the second greatest variance on the second coordinate, and so on [7].

The principal components of a collection of points in a real coordinate space are a sequence of p unit vectors, where the i-th vector is the direction of a line that best fits the data while being orthogonal to the first i-1 vectors. Here, a best-fitting line is defined as one that minimizes the average squared distance from the points to the line. These directions constitute an orthonormal basis in which different individual dimensions of the data are linearly uncorrelated. PCA is the process of computing the principal components and using them to perform a change of basis on the data, sometimes using only the first few principal components and ignoring the rest.

PCA is used in exploratory data analysis and for making predictive models (see Figure 2). It is commonly used for dimensionality reduction by projecting each data point onto only the first few principal components to obtain lower-dimensional data while preserving as much of the data's variation as possible. The first principal component can equivalently be defined as a direction that maximizes the variance of the projected data. The i-th principal component can be taken as a direction orthogonal to the first i-1 principal components that maximizes the variance of the projected data.

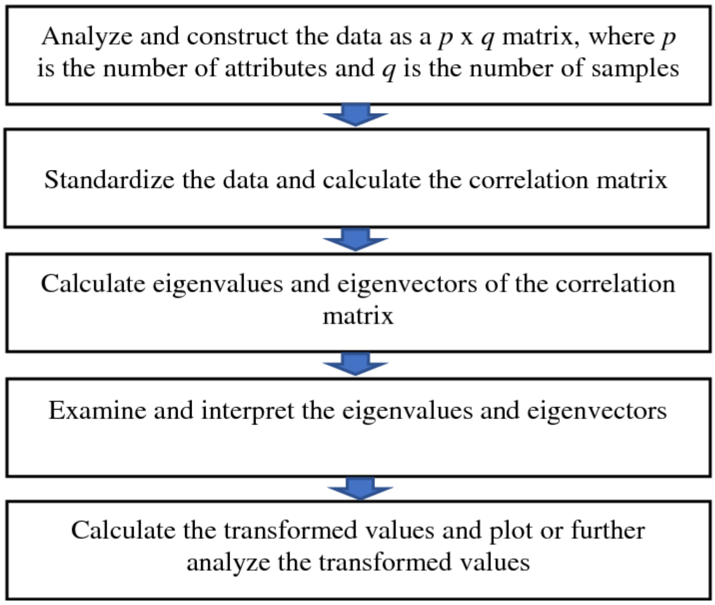

Analyze and construct the data as a p x q matrix, where p is the number of attributes and q is the number of samples
Standardize the data and calculate the correlation matrix
Calculate eigenvalues and eigenvectors of the correlation matrix
Examine and interpret the eigenvalues and eigenvectors
Calculate the transformed values and plot or further analyze the transformed values

Figure 2. PCA step.

2.3.3. Time Series Forecasting with VAR and LSTM

Vector auto regressive (VAR) is a multivariate time series model that can be used to forecast more than one variable collectively. It can be used in scenarios where multiple variables have a dependency on each other. In VAR modelling, each variable is modelled as a linear combination of past observations of itself and other variables. Therefore, it can be modelled as a system of equations, where each variable gets one equation that can be represented as vectors. Suppose we have a vector of time series data Yt, then a VAR model with k variables and p lags can be expressed mathematically in Equation (4) where, Yt, β0 and are k x 1 column vectors and β0, β1, β2, …, βp are k × k matrices of coefficients.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \varepsilon_t \tag{4}$$

If a time series is not stationary, it is essential to differentiate the time series before training the model and invert the predicted values to get the real forecast by the number of times differentiated [8].

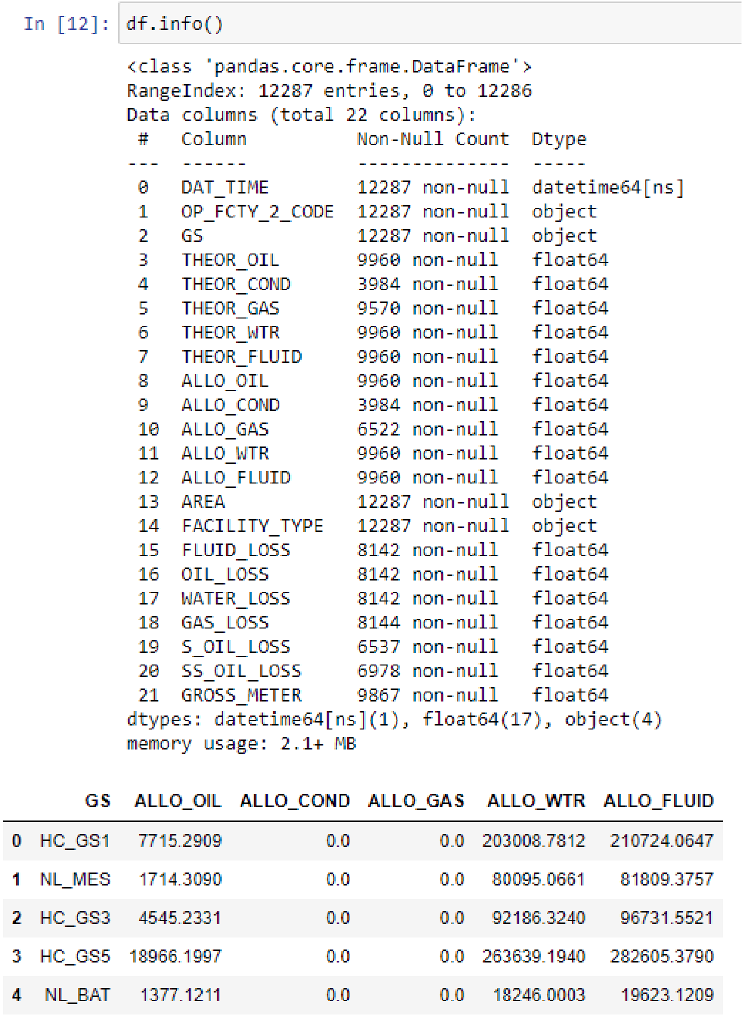
Long short-term memory (LSTM) is a special kind of recurrent neural network which makes use of sequential observations and learns from the prior memorize the sequence of information. The memorization of the prior trend of the data is done through a few gates alongside a memory line associated in an ordinary LSTM [23]. Each LSTM is a set of cells where the data streams are captured and stored. LSTMs create a transport line that connects one module to another, carrying data from the past and keeping them for the present. Using gates in each cell, data can be disposed of, filtered, or added for the next cells. Those gates are based on a sigmoidal neural network layer which can enable the cells to optionally let data pass through or discard them [24].

A sigmoid layer takes input in the range of zero and one, indicating the amount of data goes through in each cell. Estimation of zero value says nothing passes through the cell and one indicates that everything passes through the cell. There are three types of gates involved in each LSTM to control the state of each cell, forget gate, memory gate and output gate. Forget gate outputs a number between 0 and 1 to say completely ignore this and completely keep all. Memory gate chooses which new data need to be stored in the cell. Output gate decides what will be yielded out of each cell. Dissanayake et al. (2021) conclude VAR produced the best performance, followed by LSTM and ARIMA [8].

3. RESULTS AND DISCUSSION

3.1. Exploratory Data Analysis

Multiple point of production flowrate measurement is gathered through the database query. Datasets consist of multi-feature attributes, as following in Figure 3.



```
In [12]: df.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 12287 entries, 0 to 12286
Data columns (total 22 columns):
 #   Column          Non-Null Count  Dtype
---  ------          --------------  -----
 0   DAT_TIME        12287 non-null  datetime64[ns]
 1   OP_FCTY_2_CODE  12287 non-null  object
 2   GS              12287 non-null  object
 3   THEOR_OIL       9960 non-null   float64
 4   THEOR_COND      3984 non-null   float64
 5   THEOR_GAS       9570 non-null   float64
 6   THEOR_WTR       9960 non-null   float64
 7   THEOR_FLUID     9960 non-null   float64
 8   ALLO_OIL        9960 non-null   float64
 9   ALLO_COND       3984 non-null   float64
 10  ALLO_GAS        6522 non-null   float64
 11  ALLO_WTR        9960 non-null   float64
 12  ALLO_FLUID      9960 non-null   float64
 13  AREA            12287 non-null  object
 14  FACILITY_TYPE   12287 non-null  object
 15  FLUID_LOSS      8142 non-null   float64
 16  OIL_LOSS        8142 non-null   float64
 17  WATER_LOSS      8142 non-null   float64
 18  GAS_LOSS        8144 non-null   float64
 19  S_OIL_LOSS      6537 non-null   float64
 20  SS_OIL_LOSS     6978 non-null   float64
 21  GROSS_METER     9867 non-null   float64
dtypes: datetime64[ns](1), float64(17), object(4)
memory usage: 2.1+ MB
```

	GS	ALLO_OIL	ALLO_COND	ALLO_GAS	ALLO_WTR	ALLO_FLUID
0	HC_GS1	7715.2909	0.0	0.0	203008.7812	210724.0647
1	NL_MES	1714.3090	0.0	0.0	80095.0661	81809.3757
2	HC_GS3	4545.2331	0.0	0.0	92186.3240	96731.5521
3	HC_GS5	18966.1997	0.0	0.0	263639.1940	282605.3790
4	NL_BAT	1377.1211	0.0	0.0	18246.0003	19623.1209

Figure 3. Dataset head and information.

Some not a number data filled with 0 and the attributes that being dropped are, all THEOR_ and LOSS value that possible to be utilized in future research, related to production loss.

Some of informative graph are plotted to see a brief profile of the dataset. It is clearly show within the density plot in Figure 4 that the majority of the production facilities are processing oil and water, with less production of condensate and gas.

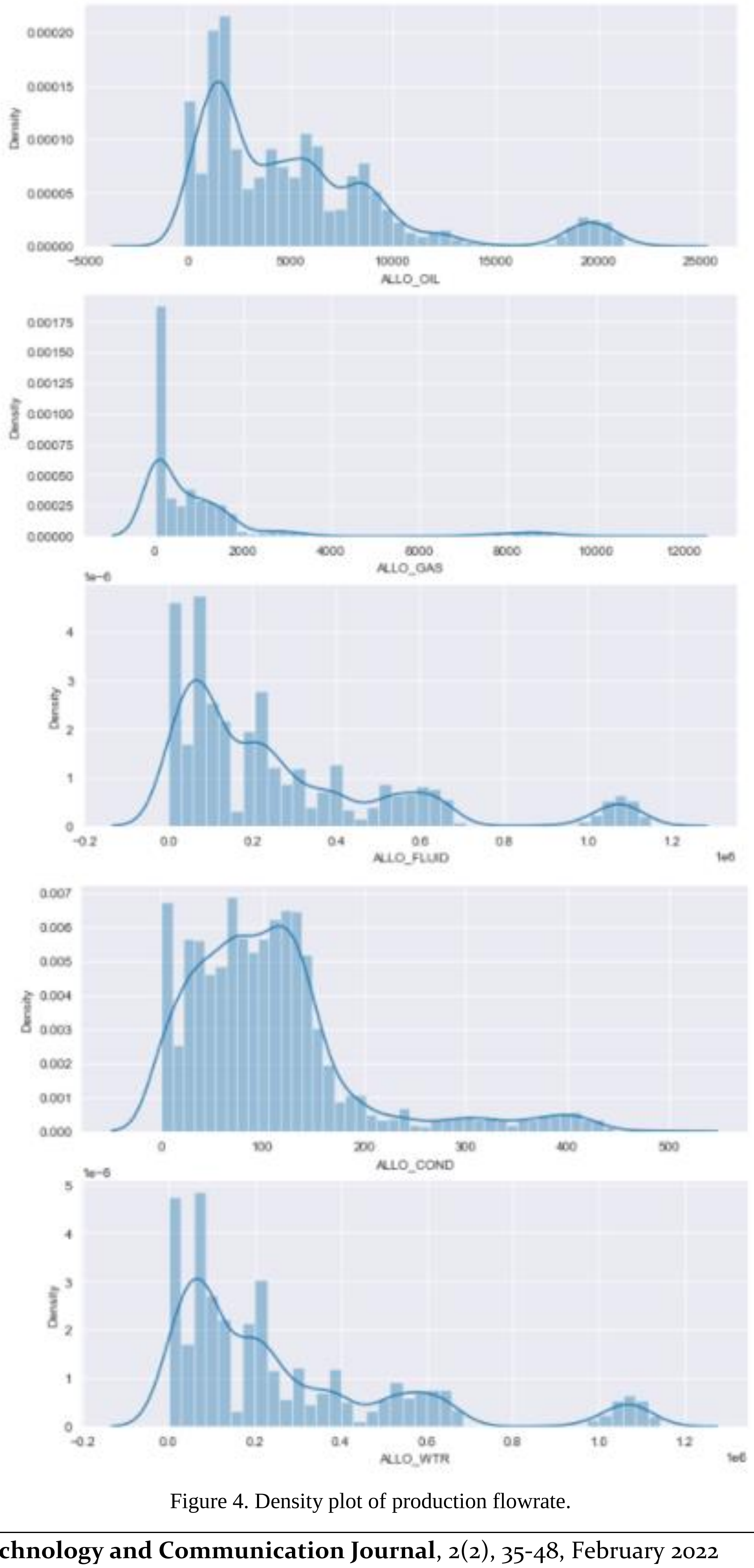


Figure 4. Density plot of production flowrate.

The heatmap correlation in Figure 5 shows that there is strong correlation between fluid and water, whereas it can be concluded that the majority of the field are having high water cut. The next strong correlation is between condensate and gas with 0.59. It is obvious that gas fields often followed by condensate production.

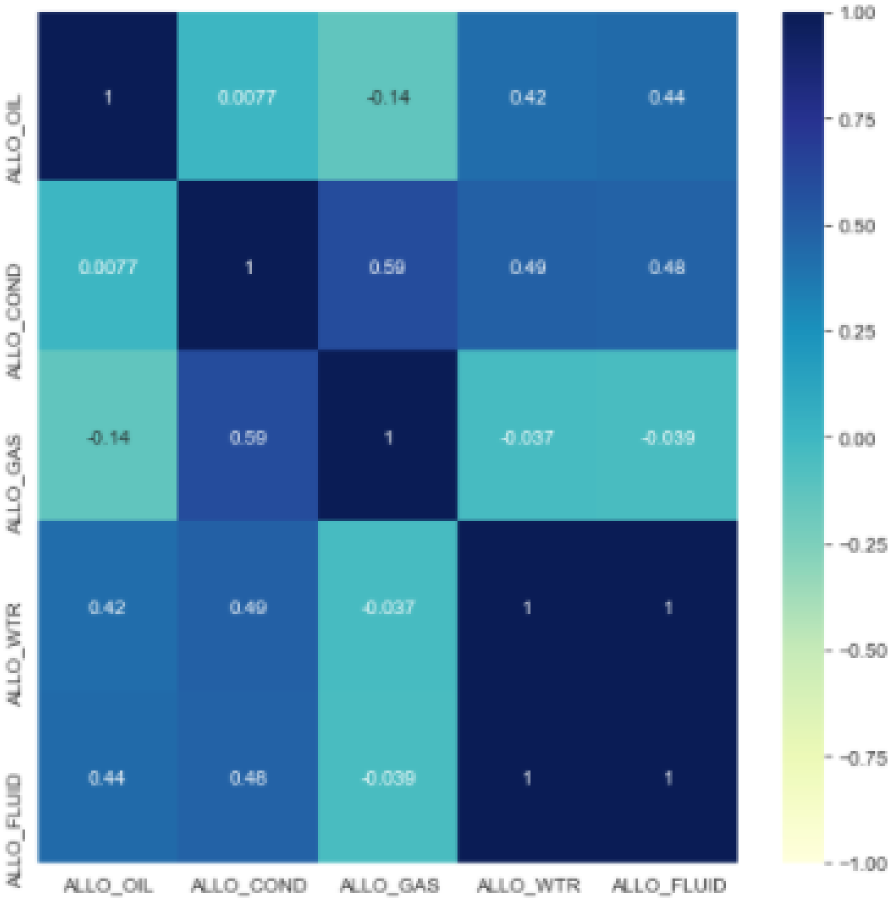


Figure 5. Correlation heatmap of measurement.

The last interesting correlation is oil with fluid and water (0.42 – 0.44), and condensate with fluid and water (0.46 – 0.49). The more fluid and water process in the production facility, it is most likely the more oil production will be gathered.

The least correlated attributes are gas and oil with -0.14. It can be interpreted as the gas fields are separated and not correlated with oil fields, or the oil fields are producing few associated gas.

3.2. K-Means Clustering and PCA Result

There are more than 30 gathering station facilities inside the dataset with various production performance. However better clustering with basis of production performance can be constructed, in order to easily capture the priority and distinguish the operating performance.

A fundamental step for any unsupervised algorithm is to determine the optimal number of clusters into which the data may be clustered. The Elbow Method is one of the most popular methods to determine this optimal value of k.

Distortion is calculated as the average of the squared distances from the cluster centers of the respective clusters. Typically, the Euclidean distance metric is used, where Inertia is the sum of squared distances of samples to their closest cluster center.

Elbow method is used to see the elbow point as the basis to decide how much n-cluster. The result show in Figure 6**Error! Not a valid bookmark self-reference.** that 4 cluster is quite represent the group of production facility cluster.

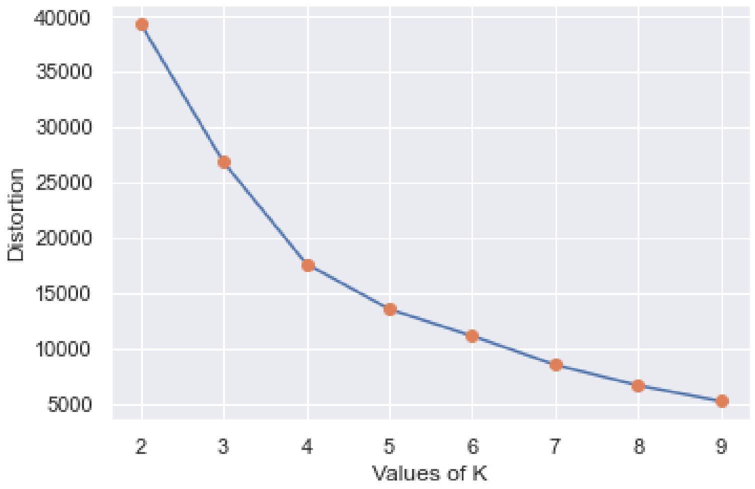


Figure 6. Elbow chart.

The following figure are the comparison of pair scatterplot of the attributes before and after clustering. The Figure 7 and 8 show cluster color of individual gathering station facilities, and the show the 4 new clusters.

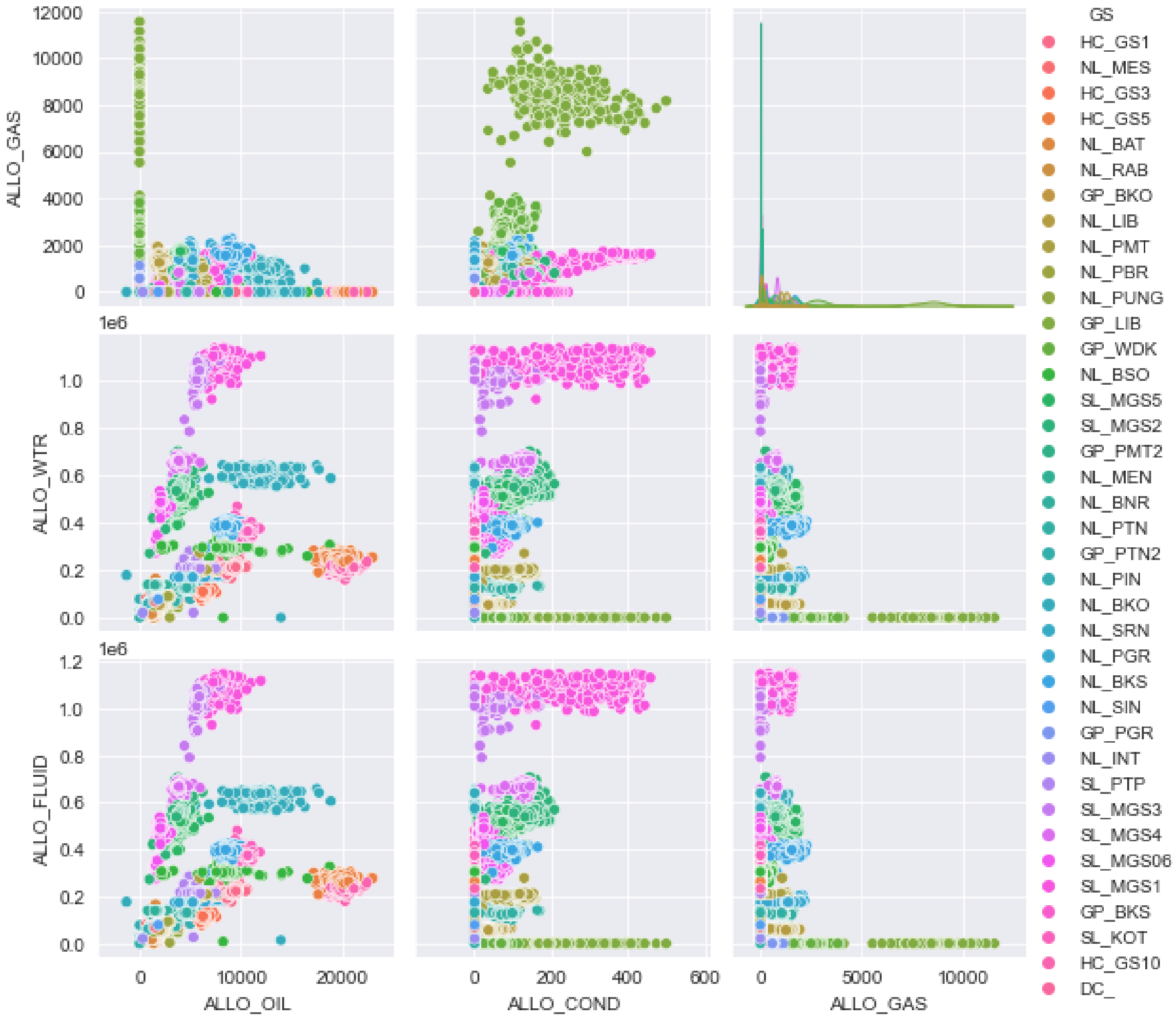


Figure 7. Scatter pair plot with gathering station clustering.

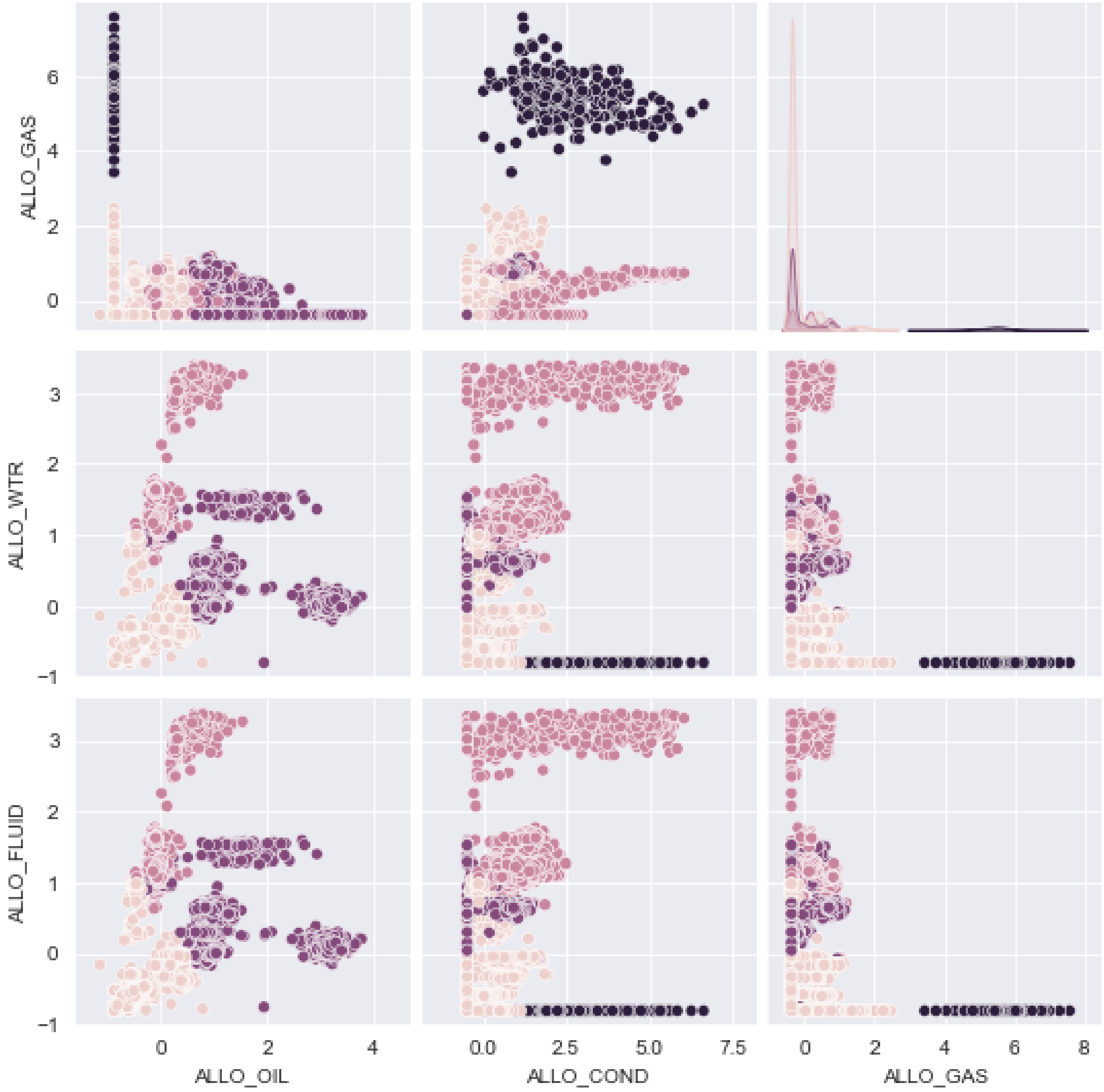


Figure 8. Scatter pair plot after K-means clustering.

The next following step is to reduce the attribute dimension using PCA, and collect the information of membership of the cluster. In order to have better view of the clustering in 3D plane, PCA 3 components are being executed. Table 1 show the head data of pca and cluster.

Standard scalar data frame is also being implemented to reduce dimension of the data. Along with the 3D plot, the membership of the 4 clusters is also being listed out, to see the characteristic of the gathering station.

Table 1. PCA 3 components.

	PCA1	PCA2	PCA3	Cluster
0	0.460837	-0.664987	1.237369	2
1	-1.197271	-0.189224	-0.284277	0
2	-0.942922	-0.425845	0.016770	0
3	1.521008	-1.621967	2.403814	2
4	-1.464446	-0.081401	-0.176032	0

The clearest view is the cluster-3 (red in 2D plot), its outlier characteristics is clearly seen in Figure 9. It is a gas plant, with only processing gas and condensate. Cluster-2 filled with high oil producer, and Cluster-1 is high water processing, whereas the cluster 0 are the lowest performer in term of hydrocarbon processing.This clustering is useful for business-oriented purpose in term of activity focus and priority, also an opportunity to develop efficient organization capability [25].

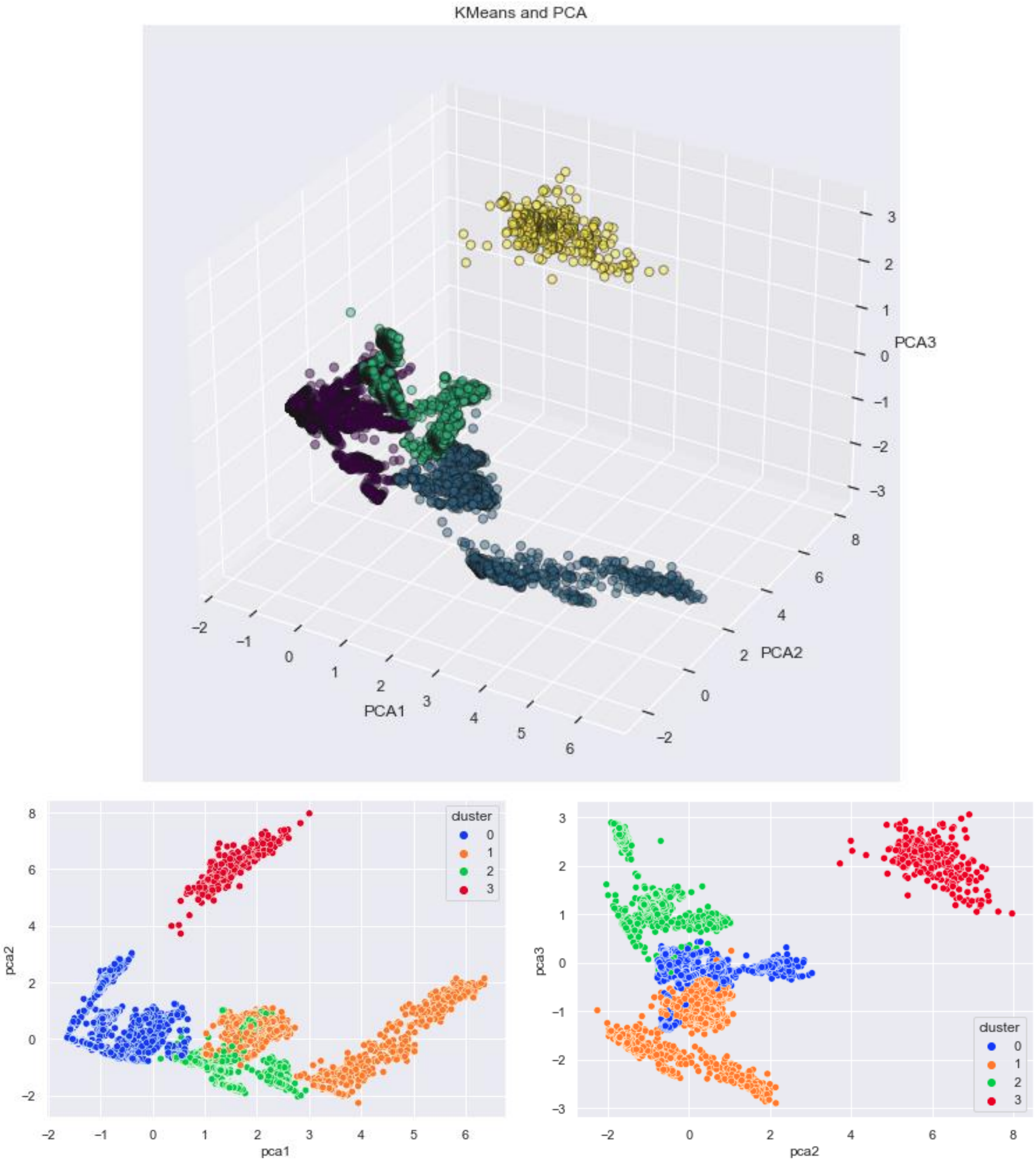


Figure 9. Plot of PCA with K-means clustering.

3.3. Time-Series Forecasting Result

Forecasting the hydrocarbon is part of the business planning projection and day to day operations as well (see Figure 10). The acceptable result of forecasting, usually being used as basis for cross-function study, budgetary, critical decision, and future growth. Figure 11 shows sample of one Gathering Station hydrocarbon which VAR provide good prediction for day-to-day forecast, of oil, gas, and water. The blue line chart in Figure 10 is the actual data, where the black chart in Figure 11 is the VAR model.

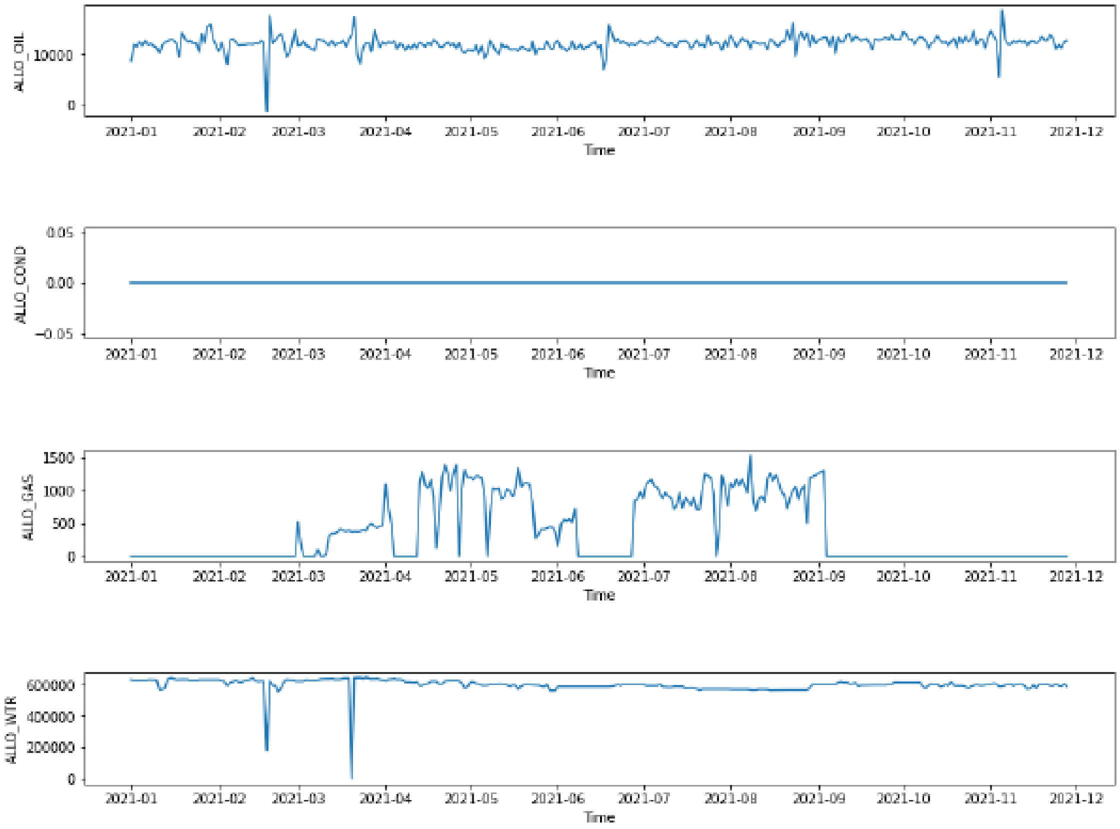


Figure 10. NL_BKO hydrocarbon fluid dataset.

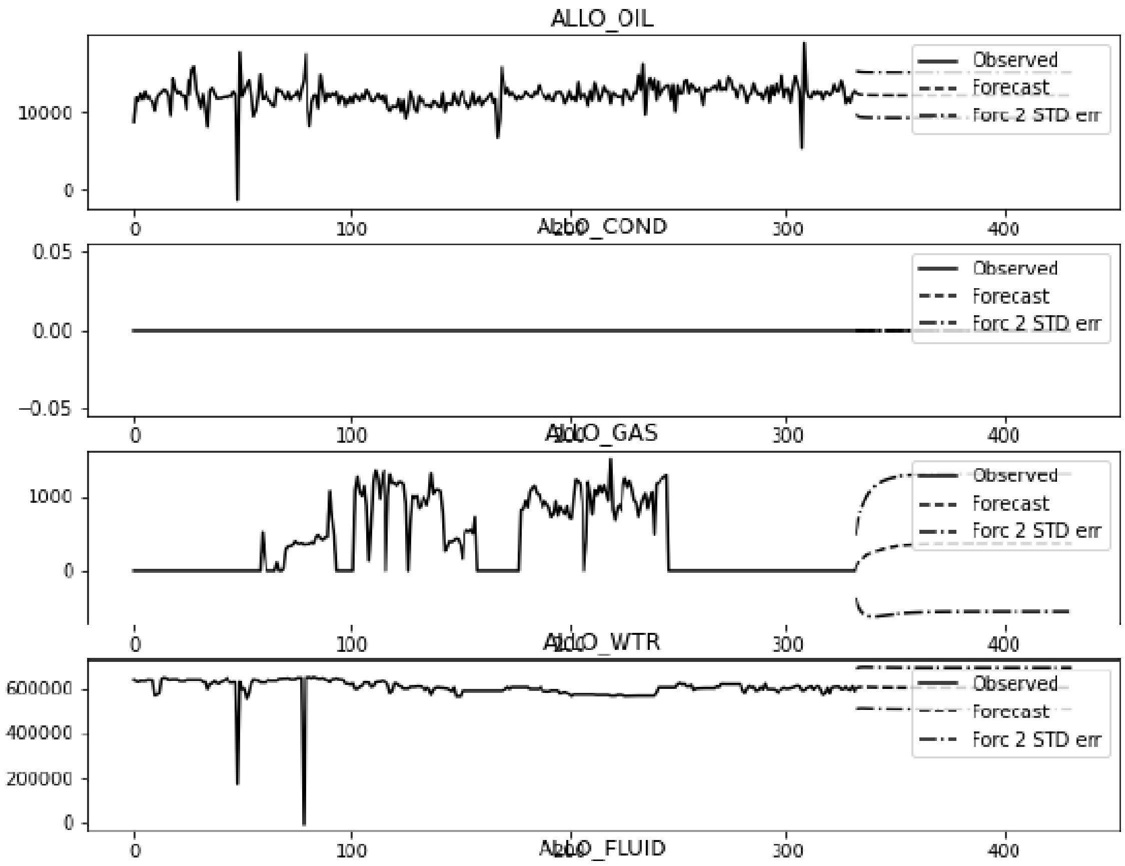


Figure 11. VAR model of NL_BKO production.

Practically, the model can be used for day-to-day operational, the more advance model of could be seen in Figure 12, LSTM provide good result visually.

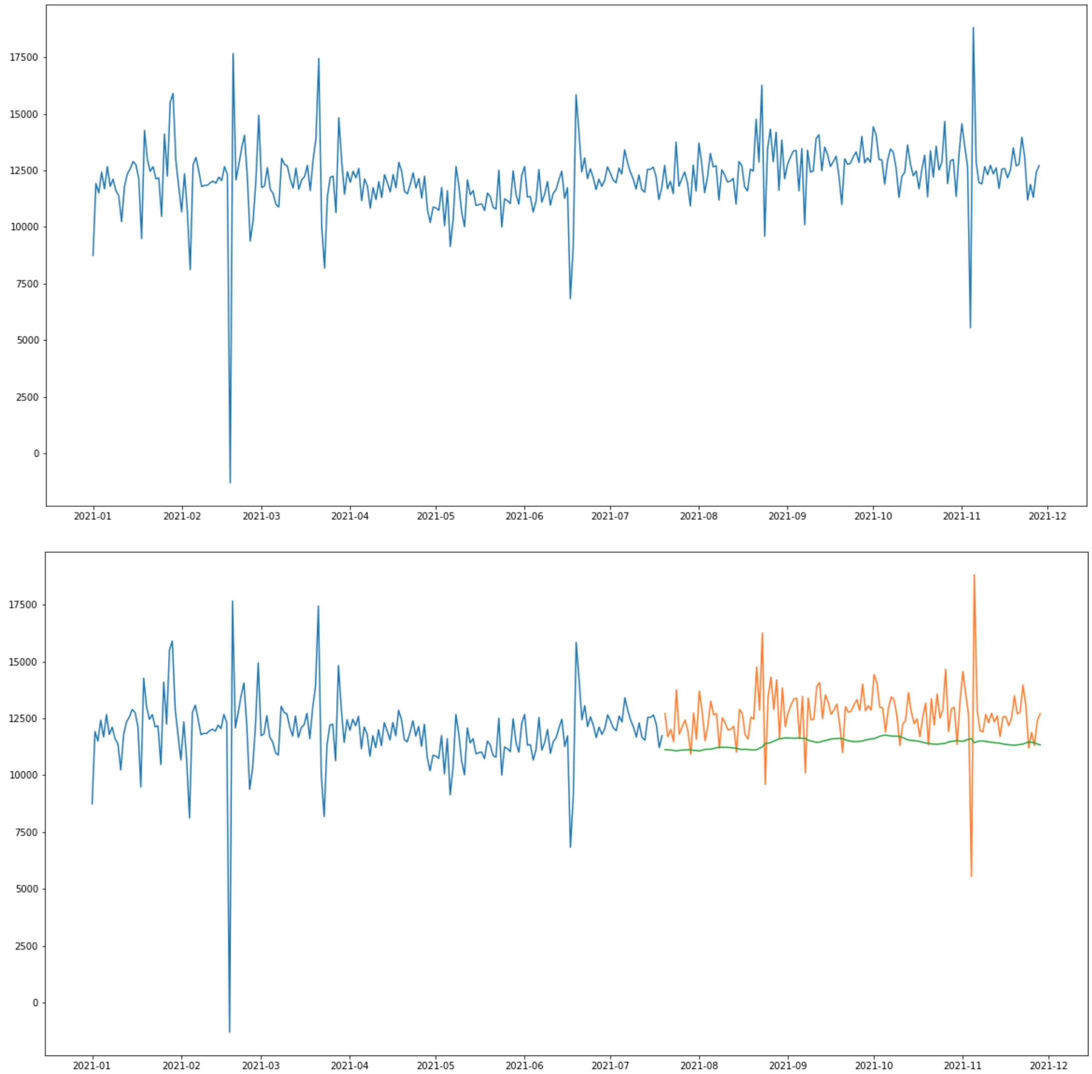


Figure 12. Oil production time-series forecast with LSTM.

Through root mean squared error (RMSE) comparison, it is clearly seen that LSTM with RMSE 1369.19 is better than VAR with enormous value 29249.59 for day-to-day operational forecasting (see Figure 13).

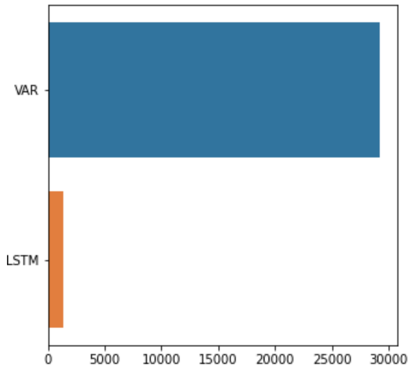


Figure 13. RMSE comparison of VAR and LSTM.

4. CONCLUSION

In a large oilfield, with many scattered production facilities, with limited resources, it is urgently required to have an integrated mapping of the overall performance. Clustering with K-means and PCA can be a good solution to provide objective grouping, to see the high-performance facility compare to less productive one, and open up the opportunity for resource allocation and prioritization. Forecasting the hydrocarbon for day-to-day decision support can be an addition to the real-time measurement. In practical, prediction and actual real time value can be compared directly to see the accuracy of the model. There is no need for model accuracy, at least it can show the trend. The forecasting model also can be as a basis for further growth case, business plan purpose, and cross-function study as well. Based on RMSE result the LSTM is better for forecasting compare to VAR. The dataset consists of loss production value, that could be an interested topic for further model and forecast.

ACKNOWLEDGMENTS

The authors would like to thanks to Politeknik Caltex Riau lecturers, and class 20MTTKB colleagues of Applied Master of Computer Engineering.

REFERENCES

[1] Woodman, M. R., Rodriguez, J., Wade, K. C., & Samsatli, N. J. (2017). New Integrated Technology for Full Production and Facilities Modelling and Optimisation. *SPE Symposium: Production Enhancement and Cost Optimisation*, D021S006R002).

[2] Hein, F. J. (2017). Geology of bitumen and heavy oil: An overview. *Journal of Petroleum Science and Engineering*, **154**, 551–563.

[3] D. Atoufi, H. & Lampert, D. J. (2020). Impacts of oil and gas production on contaminant levels in sediments. *Current Pollution Reports*, **6**, 43–53.

[4] Rimtip, N. & Tanko, N. (2021). Formulation of oil based mud from *Ricinus communis*. *ATBU Journal of Science, Technology and Education*, **9**(3), 141–153.

[5] Seyyedattar, M., Zendehboudi, S., & Butt, S. (2020). Technical and non-technical challenges of development of offshore petroleum reservoirs: Characterization and production. *Natural Resources Research*, **29**(3), 2147–2189.

[6] Kushwaha, S., Bahl, S., Bagha, A. K., Parmar, K. S., Javaid, M., Haleem, A., & Singh, R. P. (2020). Significant applications of machine learning for COVID-19 pandemic. *Journal of Industrial Integration and Management*, **5**(4), 453–479.

[7] Bro, R. & Smilde, A. K. (2014). Principal component analysis. *Analytical methods*, **6**(9), 2812–2831.

[8] Dissanayake, B., Hemachandra, O., Lakshitha, N., Haputhanthri, D., & Wijayasiri, A. (2021). A comparison of ARIMAX, VAR and LSTM on multivariate short-term traffic volume forecasting. *Conference of Open Innovations Association, FRUCT*, (28), 564–570.

[9] Harvey, A. & Kattuman, P. (2021). A farewell to R: Time-series models for tracking and forecasting epidemics. *Journal of the Royal Society Interface*, **18**(182), 20210179.

[10] Sheremetov, L. B., González-Sánchez, A., López-Yáñez, I., & Ponomarev, A. V. (2013). Time series forecasting: applications to the upstream oil and gas supply chain. *IFAC Proceedings Volumes*, **46**(9), 957–962.

[11] Anderson, R. N., Xie, B., Wu, L., Kressner, A. A., Frantz Jr, J. H., Ockree, M. A., & Brown, K. G. (2016). Petroleum analytics learning machine to forecast production in the wet gas marcellus shale. *Unconventional Resources Technology Conference, San Antonio, Texas, 1-3 August 2016*, 132–147.

[12] Wang, T., Li, T., Xia, Y., Zhang, Z., & Jin, S. (2017). Risk assessment and online forewarning of oil & gas storage and transportation facilities based on data mining. *Procedia Computer Science*, **112**, 1945–1953.

[13] Fan, D., Sun, H., Yao, J., Zhang, K., Yan, X., & Sun, Z. (2021). Well production forecasting based on ARIMA-LSTM model considering manual operations. *Energy*, **220**, 119708.

[14] Wang, Y., Li, J., Ruijie, Z., Gao, S., & Ren, Y. (2021). Realization of Accurate Modeling and Simulation of Oilfield Water Mixing Gathering and Transportation System. *Journal of Physics: Conference Series*, **1894**(1), 012101.

[15] Elijah, O., Ling, P. A., Rahim, S. K. A., Geok, T. K., Arsad, A., Kadir, E. A., Abdurrahman, M., Junin, R., Agi, A., & Abdulfatah, M. Y. (2021). A survey on industry 4.0 for the oil and gas industry: upstream sector. *IEEE Access*, **9**, 144438–144468.

[16] Alexandrov, A., Benidis, K., Bohlke-Schneider, M., Flunkert, V., Gasthaus, J., Januschowski, T., Maddix, D. C., Rangapuram, S., Salinas, D., Schulz, J. and Stella, L., Ali Caner Türkmen, A. C., & Wang, Y. (2020). Gluonts: Probabilistic and neural time series modeling in python. *Journal of Machine Learning Research*, **21**(116), 1–6.

[17] Bagheri, M., Roshandel, R., & Shayegan, J. (2018). Optimal selection of an integrated produced water treatment system in the upstream of oil industry. *Process Safety and Environmental Protection*, **117**, 67–81.

[18] Rodrigues, H. W. L., Prata, B. D. A., & Bonates, T. O. (2016). Integrated optimization model for location and sizing of offshore platforms and location of oil wells. *Journal of Petroleum Science and Engineering*, **145**, 734–741.

[19] Hollingsworth, J. L. (2015). Digital oilfield standards update. *SPE Digital Energy Conference and Exhibition*, D031S021R003.

[20] Bruno, S., De Fino, M., & Fatiguso, F. (2018). Historic Building Information Modelling: performance assessment for diagnosis-aided information modelling and management. *Automation in Construction*, **86**, 256–276.

[21] Wu, W. & Peng, M. (2017). A data mining approach combining K-means clustering with bagging neural network for short-term wind power forecasting. *IEEE Internet of Things Journal*, **4**(4), 979–986.

[22] Theocharides, S., Makrides, G., Livera, A., Theristis, M., Kaimakis, P., & Georghiou, G. E. (2020). Day-ahead photovoltaic power production forecasting methodology based on machine learning and statistical post-processing. *Applied Energy*, **268**, 115023.

[23] Sharma, A., Tiwari, P., Gupta, A., & Garg, P. (2021). Use of LSTM and ARIMAX algorithms to analyze impact of sentiment analysis in stock market prediction. *Intelligent Data Communication Technologies and Internet of Things: Proceedings of ICICI 2020*, 377-394.

[24] Liu, W., Liu, W. D., & Gu, J. (2020). Forecasting oil production using ensemble empirical model decomposition based long short-term memory neural network. *Journal of Petroleum Science and Engineering*, **189**, 107013.

[25] Cremaschi, S., Shin, J., & Subramani, H. J. (2015). Data clustering for model-prediction discrepancy reduction–A case study of solids transport in oil/gas pipelines. *Computers and Chemical Engineering*, **81**, 355–363.