CHEMICAL ENGINEERING TRANSACTIONS
VOL. 99, 2023
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Sauro Pierucci, Flavio Manenti
Copyright © 2023, AIDIC Servizi S.r.l.
ISBN 978-88-95608-98-3; **ISSN** 2283-9216
DOI: 10.3303/CET2399101

Statistical Analysis in Database of Offshore Naturally Flowing Wells with Abnormal Events

Raul M. F. U. Foronda[a], Victoria M. Fracassio[a], Rejane B. Santos[a,b], Brunno F. dos Santos[a,*]

[a]Department of Chemical and Materials Engineering (DEQM), Pontifical Catholic University of Rio de Janeiro (PUC-Rio). Rua Marquês de São Vicente, 225 – Gávea, Rio de Janeiro – RJ, 22453-900, Brazil.
[b]Federal Institute of Sul de Minas Gerais (IFSULDEMINAS), Av. Maria da Conceição Santos, 900 - Parque Real, Pouso Alegre - MG, 37550-000, Brazil.
bsantos@puc-rio.br

The development of statistical analysis in the oil and gas industry database represents the importance of seeking improvements in safety and preventing undesirable events. For accident prevention, which may cause damage to the environment and financial losses, fault detection is essential. Modern techniques on the data-driven, such as data analysis and artificial intelligence are currently considered the best options for this purpose. Regardless of advances in techniques for database manipulation, scientists spend an amount of time working on data quality improvement. Thus, the high dimensionality of data introduces computational and statistical challenges such as acquisition, treatment, processing and interpretation of data. In this paper, we used the 3W public dataset published by Vargas et al. (2019), provided by Petróleo Brasileiro S.A. (Petrobras), which contains 8 variables such as pressure, temperature and flow rate in the process of offshore natural flow wells. The database was evaluated by data exploration and transformation that enables statistical analysis. The relationship between the variables was verified using histograms, Spearman's correlation and Principal Component Analysis (PCA). The results showed that at least one variable should be removed and others should be filled in order to complete the database. The analysis, also, revealed that the variables do not follow a normal distribution, and the variables importance rank. Thereby, it was possible to reach a database with useful format.

1. Introduction

The production of oil and gas (O&G) concerns chemical and mechanical process that affect well drilling and operation. It is of common knowledge that studying the prevention of undesirable events (system failures) can avoid economic and human losses in modern engineering. From 1970 to 2016, there have been recorded more than 1,000 severe accidents in the O&G industry, which leads to uncountable casualties in addition to serious damage (Mignan at al. 2022). Modern technologies such as artificial intelligence (AI), along with cloud computing and data analysis, are currently considered the best shot to perform fault detection and process optimization (Marins et al. (2021), Sircar et al. (2021), Verheyleweghen and Jäschke (2018), Dhaif et al. (2021)). Machine Learning (ML) has gained considerable space into chemical engineering over the past decade, given the recent digital transformation that allowed the usage for newer technologies, cheapening the employment of modern frameworks to perform large resolutions with simple computational languages (e.g., scikit-learn for python) (Sircar et al. (2021), Schweidtmann et al. (2021)).
First step to develop models of predicting fault (detection and diagnosis) is to ensure the quality of data (integration, reducing duplicates and missing data). Some studies were developed by several scientists in last years, as seen in Azeroual at al. (2022) and Rahman (2023). Data scientists face challenges of analysing a large amount of data (such as big data) in chemical or petrochemical industries, due to very high-dimensional or unstructured data. The data pre-processing takes time to prepare a useful dataset, because there are quality issues, such as inconsistent values.

Paper Received: 14 December 2022; Revised: 4 March 2023; Accepted: 7 May 2023
Please cite this article as: Foronda R.M.F.U., Fracassio V.M., Santos R.B., Santos B.F., 2023, Statistical Analysis in Database of Offshore Naturally Flowing Wells with Abnormal Events, Chemical Engineering Transactions, 99, 601-606 DOI:10.3303/CET2399101

For this reason, the database community has developed numerous techniques for improving the situation (Kendel at al. (2011)), just as functions in programming languages (Python, R and others).Statistical analysis is a great tool to evaluate the relationship between variables. Classical correlation coefficient, as Spearman matrix, are widespread used, seen in Li et al. (2021) and Hua et al. (2020). The Principal Component Analysis (PCA) is an effective technique for discovering the hidden feature from variables (Li and Huang (2020), Mei et al. (2017)).This study aimed to initiate a more detailed analysis of dataset sensed by Vargas et al. (2019), with the intention of detecting failures caused in the process of offshore natural flow wells. Therefore, the subdataset chosen, from 3W dataset, was evaluated for cleaning and manipulation data. The relationship between the variables was verified using Spearman's correlation and a PCA analysis.

2. Materials and Methods

For this work, the database produced by Vargas et al. (2019), 3W dataset, was assessed. The 3W dataset is composed of three types of instances, real data that occurred in wells produced by Petrobras, simulated and hand-drawn. However, in this work only real data was used. Each instance of the 3W database is composed of eight variables, as shown in Table 1, from sensors of oil production systems. Besides the eight variables, the instances have an additional variable that is a vector called a class, which establishes three situations of the type: normal, transient anomaly and stable state of anomaly.

Table 1: Variables contained in the 3W dataset, adapted from Vargas et al. (2019).

Variable	Description	Units
P-PDG	Pressure in the fluid at the Permanent Downhole Gauge (PDG)	Pa
P-TPT	Pressure in the fluid at the Temperature and Pressure Transducer (TPT)	Pa
T-TPT	Temperature in the fluid at the Temperature and Pressure Transducer (TPT)	°C
P-MON-CKP	Upstream fluid pressure to the valve Production Choke (PCK)	Pa
T-JUS-CKP	Fluid temperature downstream of the valve Production Choke (PCK)	°C
P-JUS-CKGL	Fluid pressure downstream of the control valve by gas lift	Pa
T-JUS – CKGL	Fluid temperature downstream of the gas lift control valve	°C
QGL	Gas lift flow	m^3/s
Class	Indicates the status of each anomaly along the time series: normal period, transient anomaly and anomaly steady state	

In this work, the flow instability subdataset was adopted, due it's acceptable ranges of variables, as referred by Vargas et al. (2019). This subdataset contains 2,462,076 information vectors for each variable.

2.1 Software tool

All analysis was conducted on Google Colaboratory (Colab), Python 3.7.13, available for free as web application. The libraries used were *pandas* (Pandas, 2023), *NumPy* (Numpy, 2022), matplotlib 3.7.0 (Matplotlib, 2023), seaborn (Seaborn, 2022) and scikit-learn (Scilit-learn, 2023).

2.2 Data analysis and wrangling

The data underwent information counting on each variable to check for missing data and some basic statistics, such as mean, median, standard deviation, i.e. the functions ".isnull()", "sum()" and "describe()" were applied. And to fill missing data it was used “.fillna(method=’bfill’)”, this methodology propagates the first observed non-null value backward until another non-value is met.

2.3 Statistical analysis

After adjustments, the database was evaluated for statistical analysis. First, histograms (“.hist()”) were performed to qualitatively verify the frequency distributions.
Correlation coefficient can describe the monotonicity of two variables. The Spearman coefficient is a correlation measure with strictly monotone transformation invariance. The heat map can express correlation diagram, +1 or -1 indicating how much stronger the relationship between variables. The “.corr()” and “.heatmap()” functions were used to obtain values corresponding to correlation matrix.
The strength of a Spearman correlation is categorized using the following guidelines for the absolute value of coefficient (Rs), Zhao et al. (2022):

$$0.80 \leq |Rs| \leq 1.0 \text{ "}very\ strong\text{"};$$
$$0.60 \leq |Rs| \leq 0.79 \text{ "}strong\text{"};$$
$$0.40 \leq |Rs| \leq 0.59 \text{ "}moderate\text{"};$$
$$0.20 \leq |Rs| \leq 0.39 \text{ "}weak\text{"};$$
$$0.00 \leq |Rs| \leq 0.19 \text{ very weak.}$$

Principal Component Analysis (PCA) was adopted for dimensionality reduction or to arrange the importance of variables by rank. Then, the main objective was identifying the inter-relationship among the independent variables. The “PCA()” function along with graphical techniques based on *scree plot* (which shows the eigenvalues in decreasing order) were analysed.

3. Results and Discussion

The results of previous statistical analysis of the subdataset *flow instability* (from 3W dataset) showed that all variables had missing data (Not a Number - NaN), such as: P-PDG - 750.0; P-TPT - 1,067.0; T-TPT - 1,056.0; P-MON-CKP - 672; T-JUS-CKP - 1,126; P-JUS-CKGL - 1,243,569; T-JUS-CKGL - 2,462,076; QGL - 579,609. The percentage of missing data were calculated, it is possible to assess the impact in database. Thereby, the variables with the highest number of missing data were P-JUS-CKGL and T-JUS -CKGL, 50.50% and 100%, respectively. This fact had repercussions on the removal of these two variables from the database.
The behaviour of the variables was evaluated, displayed in Figure 1. Each variable was portrayed as completely different, considering the data collected for seven months. The database revealed a type of event in oil wells that underwent relevant changes, but with tolerable amplitudes.

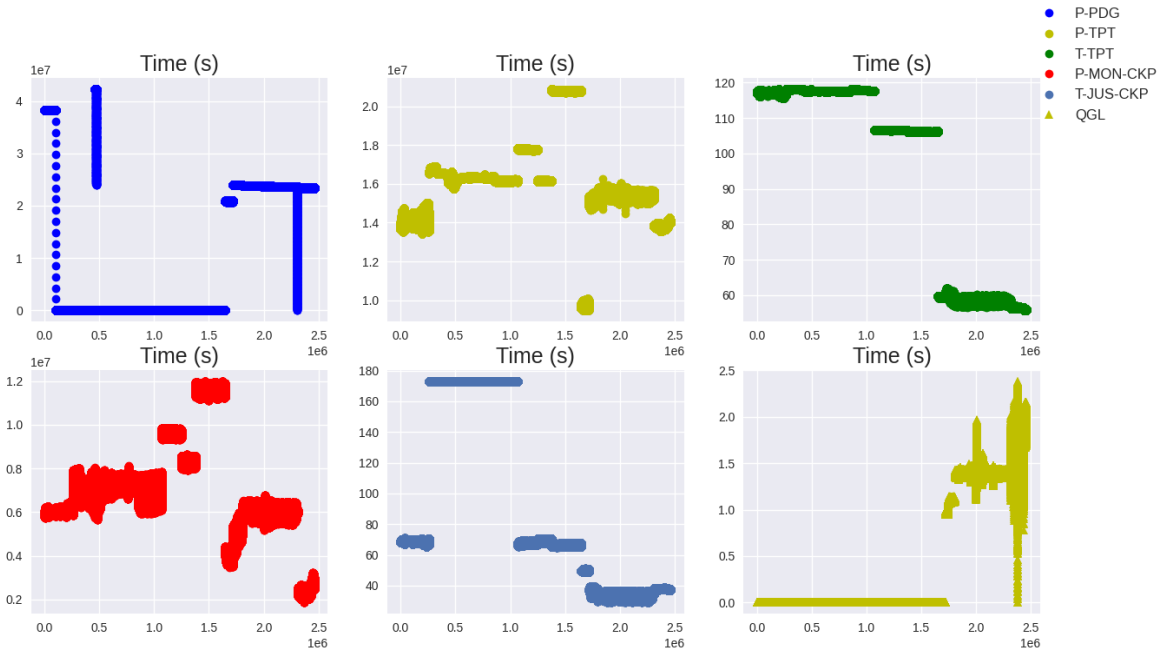

Figure 1: Behaviour of the variables P-PDG, P-TPT, T-TPT, P-MON-CKP, T-JUS-CKP, P-JUS-CKGL, T-JUS-CKGL and QGL raw data.

In parallel, some information on the considered variables was verified, such as arithmetic mean, standard deviation, maximum and minimum values, according to Table 2.

Table 2: Information on the variables considered.

	P-PDG	P-TPT	T-TPT	P-MON-CKP	T-JUS-CKP	QGL
Mean	9.59693E+06	1. 61094E+07	9.53257E+01	7.13235E+06	9.21587E+01	5.49359E-01
std	1.27087E+07	2.17084E+06	2.64399E+01	2.21848E+06	5.80663E+01	6.90434E-01
Min	0.00000E+00	9,47822E+06	5.57029E+01	1.85999E+06	2,90300E+01	0.00000E+00
Max	2.46207E+06	2.08888E+07	1.18199E+02	1.19983E+07	1.73096E+02	2.38415E+00

The frequency distribution for the variables was performed by histograms, Figure 3. This analysis indicates that the majority of the grouped observations has a normal (or Gaussian) distribution. This technique tests parametric procedures.

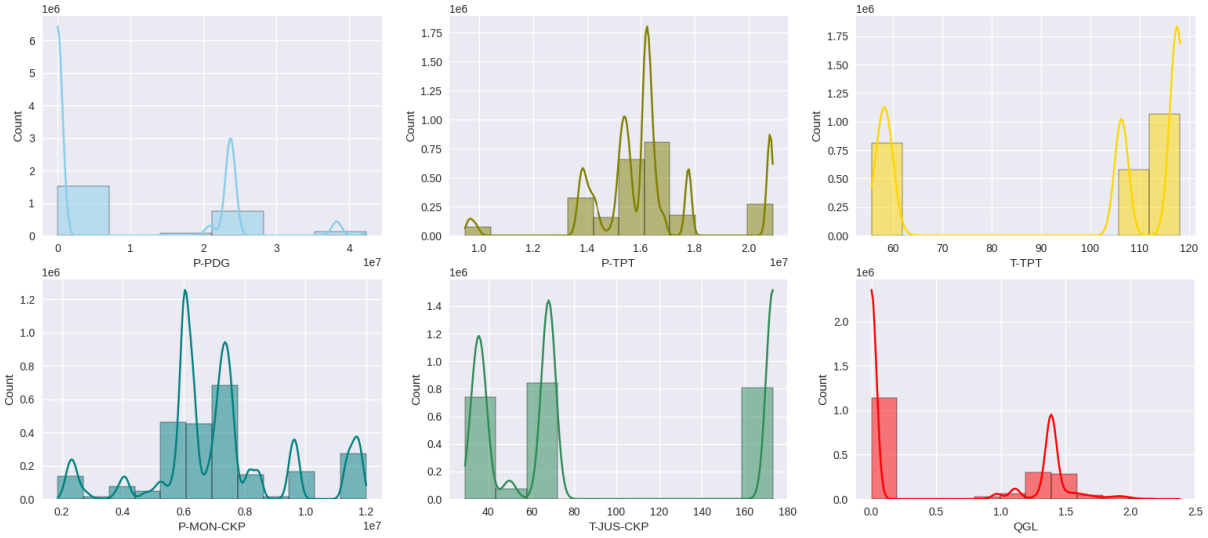

Figure 3: Histogram of variables.

As it can be seen, the variables do not follow a normal distribution. This behaviour was expected since changes are seen in different amplitudes. Other fact, the variables was collected by real scenario, which means that lot of noise is contained. With the aim of avoiding problems in statistical analysis, data manipulation for including elements in database. In many situations, the variables are filled he empty elements with the next non-empty value of the sequence, using the attribute *"bfill"* (backward fill) in *.fillna(), pandas*. The method *.describe()* was used to confirm the filling.After that, the Spearman correlation matrix was performed to investigate the relationship between two variables and to evaluate the nonlinear correlation based on non-parametric statistical. Figure 4 shows all correlations between the variables from database.

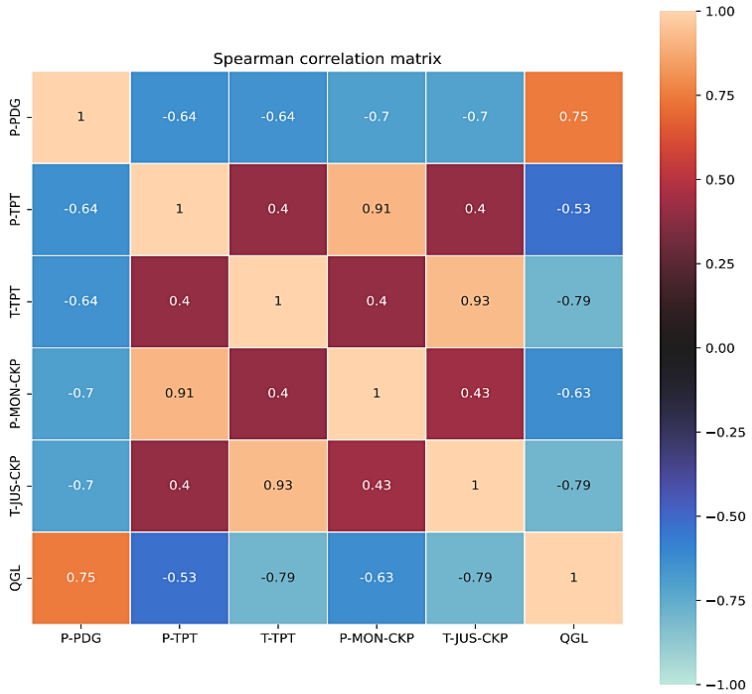

Figure 4: Histogram of variables.

According to matrix (Figure 4), the variable P-PDG has a strong correlation with all variables, in other hand, only positively with QGL, which means they grow in the same direction, observed in Figure 1.
While the variable P-TPT has a very strong correlation with the variable P-MON-CKP (positively), moderate with the variable QGL (negatively) and weak with the variables T-TPT and T-JUS-CKP. Probably, the flow instability passing through tube until platform undergoes disturbances from the environment.The variable T-TPT has a very strong correlation with the variable T-JUS-CKP (positively), strong with the variable QGL (negatively) and moderate with the variables P-TPT and P-MON-CKP (positively).The variable P-MON-CKP has a moderate correlation with the variable T-JUS-CKP (positively) and strong with the variable QGL (negatively). Finally, the variable T-JUS-CKP has a strong correlation with the variable QGL (positively).
In general, Spearman correlation matrix showed that the most of temperatures have moderate correlation with pressures, but have strong correlation with each other and vice-versa.The analysis of database by PCA has shown the values of variances for the six new principal components (PC). The PCs presented a cumulative percentage of variance, which implies that these they can be used to explain the database, indicating the order of importance. PC1 was 96.23%, PC2 99.72% and from the addition of the PC3 100% were reached. PCA transformation expresses the correlation between PCs and the original variables, and these results make it clear that most of variance is explained by PC1. Another tool, the scree plot (Figure 5) finds a sharp reduction in the size of the eigenvalues, with the rest of the smaller eigenvalues remain in plateau.

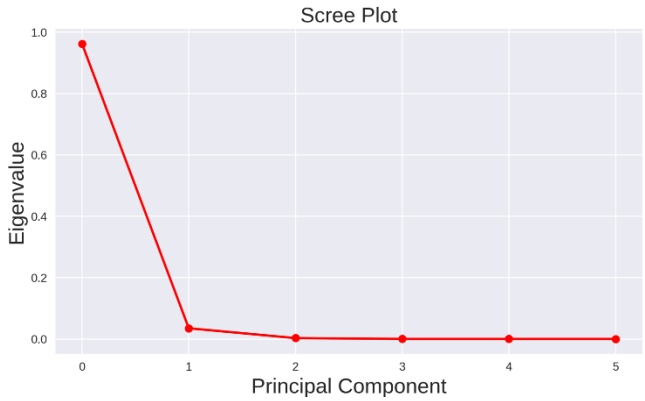

Figure 5: Scree plot with the threshold of variance explained approach.

The Figure 5 suggests that a useful model for data may have three PCs with total explainability of variables. The rank of variables calculated by PCA can be seen in Table 3. Thus, P-PDG is the most important variable, the PCA analysis indicates that it has a stronger influence according to PC1.

Table 3: Variable's rank according to PCA

	Components
P-PDG	0.989895
QGL	0.000000
T-TPT	-0.000002
T-JUS-CKP	-0.000003
P-TPT	-0.091859
P-MON-CKP	-0.108031

4. Conclusions

This paper proposed to investigate a real database, from 3W dataset, with instances of types of undesirable events that may happen in offshore naturally flowing oil and gas wells. The subdataset was chosen from the flow instability with 2,462,076 information vectors. In order to understand their statistical behaviour, it was performed some analysis. The subset reveals that variables showed missing data and different correlation each other. PCA allows identifying the variable's rank: P-PDG>QGL>T-TPT>T-JUS-CKP>P-TPT>P-MON-CKP. From these results is possible to compile a robust database and to develop several investigations, such as machine learning models. For further researches, it is recommended the implementation the same analysis in all database, and, also, to conduct research on the development of models based on artificial intelligence.

Acknowledgments

The authors would like to thank the CNPq/MCT, CAPES, FAPERJ (E-26/200.282/2023-283570) and FINEP for the financial support to the Department of Chemical and Materials Engineering (DEQM) at the Pontifical Catholic University of Rio de Janeiro (PUC-Rio).

References

Azeroual O., Schöpfel J., Ivanovic D., Nikiforova A., 2022, Combining Data Lake and Data Wrangling for Ensuring Data Quality in CRIS, 211, 3-16.

Dhaif R. A., Ibrahim A. F., Elkatatny S., Shehri D. A., 2021, Prediction of oil rates using Machine Learning for high gas oil ratio and water cut reservoirs. Flow Measurement and Instrumentation, 82, 1-9.

Hua L., Xiao F., Li Y., Huang H., Zhao K., Yu K., Hettiarachchi C., 2020, A potential damage mechanism of rubberized cement under freeze-thaw cycle, Constr. Build. Mater., 252, DOI:10.1016/j.conbuildmat.2020.119054.

Kandel S., Heer J., Plaisant C., Kennedy J., Ham F., Riche N. H., Weaver C., Lee B., Brodbeck D., Buono P., 2011, Research directions in data wrangling: Visualizations and transformations for usable and credible data, 10. DOI: 10.1177/1473871611415994

Li W., Huang Y., 2020, A combined method of cross-correlation and PCA-based outlier algorithm for detecting structural damages on a jacket oil platform under random wave excitations, Applied Ocean Research, 102. DOI: 10.1016/j.apor.2020.102301

Li Z., Gao X., Lu D., 2021, Correlation analysis and statistical assessment of early hydration characteristics and compressive strength for multi-composite cement paste, Construction and Building Materials, 310.

Marins A. M., Barros B. D, Santos I. H., Barrionuevo D. C., Vargas R. E.V., Prego T. M., Lima A. A., Campos M. L. R., Silva E. A. B., Netto S. L.,2020, Fault detection and classification in oil wells and production/service lines using random forest, Journal of Petroleum Science and Engineering, 197.

Matplotlib, 2022, Matplotlib 3.7.0 documentation < https://matplotlib.org/stable/index.html> accessed 10.02.2023

Mei C., Chen Y., Zhang H., Chen X., Liu G., 2017, Development of a multi-model strategy based soft sensor using gaussian process regression and principal component analysis in fermentation processes, Chemical Engineering Transactions, 61, 385-390 DOI:10.3303/CET1761062

Mignan A., Spada M., Burgherr P., Wang Z., Sornete D., 2020, Dynamics of severe accidents in the oil & gas energy sector derived from the authoritative ENergy-related severe accident database. PLOS ONE, 1-14.DOI: 10.1371/journal.pone.0263962

Numpy, 2023, Numpy documentation < https://numpy.org/doc/stable/index.html> accessed 10.02.2023.

Pandas, 2023, Pandas documentation < https://pandas.pydata.org/docs/index.html> accessed 10.02.2023.

Rahman A. 2023, 9 - Data collection, wrangling, and pre-processing for AI assurance, AI Assurance, 321-338. DOI: 10.1016/B978-0-32-391919-7.00022-6

Schweidtmann A. M., Esche E., Fischer A., Kloft M., Repke J., Sager S., Mitsos A.,2021, Machine Learning in Chemical Engineering: A Perspective, Chemie Ingenieur Technik, 93, 2029-2039.

Scikit-learn, 2023, scikit-learn < https://scikit-learn.org/stable/#> accessed 10.02.2023.

Seaborn, 2022, seaborn: statistical data visualization < https://seaborn.pydata.org/ > accessed 10.02.2023.

Sircar A, Yadav K., Rayavarapu K., Bist N., Oza H., 2021, Application of machine learning and artificial intelligence in oil and gas industry. Petroleum Research, 6, 379-391.

Vargas R. E. V., Munaro C. J., Ciarelli P. M., Medeiros A. G., Amaral B. G., Barrionuevo D. C., Araújo J. C. D., Ribeiro J. L., Magalhães L. P., 2019, A realistic and public dataset with rare undesirable real events in oil wells, 181.

Verheyleweghen A., Jäschke J.,2018, Oil production optimization of several wells subject to choke degradation, IFAC-PapersOnLine, 51, 1-6, DOI: 10.1016/j.ifacol.2018.06.346.

Zhao G., Ding W., Tian J., Liu J., Gu Y., Shi S., Wang R., Sun N., 2022, Spearman rank correlations analysis of the elemental, mineral concentrations, and mechanical parameters of the Lower Cambrian Niutitang shale: A case study in the Fenggang block, Northeast Guizhou Province, South China, Journal of Petroleum Science and Engineering, 208. DOI:10.1016/j.petrol.2021.109550