Piotr Łętkowski, Andrzej Gołąbek, Paweł Budak, Tadeusz Szpunar Instytut Nafty i Gazu – Państwowy Instytut Badawczy

Robert Nowak, Jarosław Arabas Politechnika Warszawska

# Factor analysis of measurement data using the example of physicochemical well data for shale formations

The paper presents the use of statistical factor analysis for the reduction of the number of physicochemical measurement variables, based on the example of well data from shale formations. The main objective of the conducted analyses was the reduction of the number of measurement variables in a manner which would enable the preservation of the possibly high amount of information about the variability of the original data. The presented procedure consists of two main stages: 1) the elimination of correlated variables, 2) the actual factor analysis. The method turned out to be effective and it can constitute a basis for further analyses, e.g. an agglomeration analysis.

Key words: shale formations, physicochemical measurements, factor analysis.

# Analiza czynnikowa danych pomiarowych na przykładzie danych fizykochemicznych odwiertów formacji łupkowych

W artykule przedstawiono zastosowanie statystycznej analizy czynnikowej do redukcji liczby fizykochemicznych zmiennych pomiarowych na przykładzie danych z odwiertów w formacjach łupkowych. Celem przeprowadzonych analiz była redukcja liczby zmiennych pomiarowych w sposób umożliwiający zachowanie możliwie dużej ilości informacji na temat zmienności pierwotnych danych. Przedstawiona procedura składa się z dwóch podstawowych etapów: 1) eliminacja zmiennych skorelowanych, 2) właściwa analiza czynnikowa. Metoda okazała się być efek-tywna i może stanowić podstawę dla dalszych analiz, np. analizy aglomeracyjnej.

Słowa kluczowe: formacje łupkowe, pomiary fizykochemiczne, analiza czynnikowa.

### Introduction

Depending on the degree of the identification of a reservoir we may deal with several to over a dozen measurement values (measurement variables) characterising the petrophysical, geochemical, lithological or mechanical properties in each well in the reservoir. As the measurements progress, for each well we get a set of measurement variables of various types (measured in different units), which are not only correlated with each other to a varying extent, but also contain a very high number of cases, meaning the individual measurements taken along the depth of the well. The high number of variables and the amount of individual measurements are the cause of redundancy, resulting in difficulties associated with their interpretation, and make the detection of the internal structure of the data more difficult.

Limiting the number of measurement variables would allow their more effective analysis and the determination of the characteristic values for each well. These values would constitute a basis for the assessment of the similarity of data originating from various wells. To this end, the use of basic statistical values such as the mean value or variance is

# NAFTA-GAZ

highly insufficient, since it does not take into account their variability, or does so to a minor extent. On the other hand, the reduction of the number of measurement variables must be connected to the loss of information, which is disadvantageous for obvious reasons.

The purpose of using the analysis presented in the paper is to limit the amount of measurement data characterising a well by eliminating the correlated values and replacing the remaining measurement variables with a smaller number of replacement variables, at the same time providing the control of the amount of lost information. The resulting replacement variables may be used, e.g., to determine the features of similarity between the wells. Searching for groups of similar wells has, e.g., significance in reservoir simulations, where the possibility to define a structure of the alternating layers of a simulation model allows the determination of the possible directions of the flow of reservoir fluids. Another possible application is the use of similarity features to predict the extraction from wells providing access to shale formations, in a situation of the lack of production test data. A statistical analysis involving the search for similar data clusters in measurement sets for shale rocks for the purpose of their segmentation is presented in [2]. Another example of the use of statistical methods for the categorisation of measurement data is [3]. A direct inspiration to conduct the following analysis was the desire to construct a method of the identification of similar wells in order to implement it in a computer system for collecting and exchanging data for shale rocks [1]. The analysis presented in the paper constitutes a first stage of such a method of identification.

The reduction of the number of variables should proceed in two stages. The first stage is the so-called "screening" of data, involving the rejection of implausible or incomplete measurement variables, which if taken into account in further analysis could cause the distortions of the results. The second stage involves replacing the original measurement variables with a smaller number of new, representative and mutually uncorrelated variables which, however, contain the fullest possible information about the relations between the original measurement variables. In this manner, for each well we get several representative measurement variables containing the majority of information about the properties which we are interested in and which could constitute a basis for further analysis.

The method suggested for the reduction of the number of variables is based on one of the statistical techniques of data analysis – the factor analysis. The statistical methods of data analysis are commonly used when analysing the correlation and structure of mutually dependent datasets. The following elements may be distinguished in the procedure of factor analysis:

- data "screening",
- the elimination of correlated variables,
- the principal component analysis.

The paper is dedicated to the use of factor analysis for the examination of the correlation and the internal structure of data based on the example of the available geochemical data from wells in shale formations. The analysis will involve the data from the ten following wells: L-1, L-2H, O-2, O-3, K-1, G-1, T-1, W-1, Z-1, B-1, with the analysis taking into account the following geochemical measurement values:

TOC - total organic carbon content [wt%],

- $T_{max}$  the temperature at which the maximum amount of hydrocarbons is created during cracking of kerogen [°C],
- S1 the amount of free hydrocarbons content present in a rock sample [mg HC/g of rock],
- S2 the amount of hydrocarbons released during the original cracking of kerogen [mg HC/g of rock],
- S3 the amount of carbon dioxide released during the destruction of organic substance [mg CO<sub>2</sub>/g of rock],
- PI the so-called generation index,
- PC-pyrolytic carbon content [wt%],
- RC- residual carbon content [wt%],
- HI hydrogen index [mg HC/g TOC],
- OI oxygen index [mg CO<sub>2</sub>/g TOC],
- MINC total mineral carbon content [wt%].

### The analysis of principal components

The analysis of principal components involves searching for straight lines which are best adjusted to the clouds of measurement points in a vector space (the space of measurement variables and measurements), in accordance with the least squares criterion. As a result, this enables the determination of the principal components (meaning the new variables known as factor variables) defining a vector subspace with a lower dimension, which reflects the original vector space. Although the first factor (factor variable) is separated in such a manner that it could explain the variance of original measurement variables to the greatest possible extent, virtually at no times does it recreate it in its entirety. Which is why this part of the variance which remains unexplained should be explained by the next (second) factor, etc. The number of factors separated in this manner does not inherently exceed the number of output variables. The purpose of the analysis is the reduction of the number of variables in such a way as to keep the largest possible part of the information about the structure of the original data. In the mathematical approach, determining the factor variables involves the symmetrical diagonalisation of the correlation matrix. The obtained result is a new set of variables (factors), which constitute linear combinations of the original measurement variables, and are also mutually uncorrelated. In this manner, a smaller number of new variables is generated, nonetheless explaining the original variability of measurement data to the greatest possible extent.

If the number of principal components (factor variables) equals the number of original variables, then such a solution inherently recreates 100% of their variances. However, in this case the purpose for which the analysis is being conducted (meaning the reduction of the number of variables) is not achieved. It is therefore necessary to adopt a number of factor variables which is smaller than the number of original measurement variables, at the expense of the variance being explained. The choice of the optimal number of factor variables remains an important problem. Two effective criteria for the selection of the number of principal components exist and are used:

The Kaiser criterion – according to this criterion one should take into account those principal components (factor variables) which have their eigenvalues higher than 1. The criterion may be interpreted in such a manner that if a factor variable does not take into account at least one original measurement variable, then it should be rejected.

**The Cattel criterion** (the so-called scree test) – the criterion recommends finding such a point on the plot of

eigenvalues, to the right of which a gentle drop in the eigenvalues values takes place. This criterion usually results in a higher number of factor variables compared to the Kaiser criterion.

It is also possible to adopt the percentage criterion. This means that we leave as many factor variables as necessary to explain the arbitrarily selected percentage of the variances of original variables.

Regardless of the criterion adopted for the selection of the number of principal components, it is not uncommon to obtain a relatively high number of factor variables as a result of conducting a factor analysis and adopting one of the above-mentioned criteria. This usually happens when a too high number of measurement variables constitutes a basis for the analysis. We can then divide the measurement variables into two groups: active variables and additional variables, and conduct the whole analysis in two stages. In the first stage we will conduct a full analysis for variables qualified to the group of active variables, and next we will determine the correlations between the resulting factor variables and additional variables. In the second stage we will repeat the analysis for active variables and those additional variables, for which correlation with the factor variables obtained in the first stage exceeded, e.g., 50%. This solution leads to the further reduction of the number of principal components, simultaneously taking into account (although in a limited scope) the full list of measurement variables.

# The elimination of correlated variables

The first stage is the screening of input data, involving the selection of wells for which the same set of data is available, and the rejection of incomplete or uncertain data (cases). Subsequently, due to the high number of variables constituting the measurement results, it is necessary to eliminate the dependent measurements. The measure of the dependence between the individual measurements for each well will be the Pearson correlation coefficient. The correlation matrices of measurement variables for three selected wells: L-1, O-3 and W-1 are presented below.

In order to eliminate the dependent variables, TOC was adopted as the reference variable, and subsequently those variables for which the coefficient of correlation with TOC was higher than 0.8 were removed from further analysis. Those variables are marked red in Tables 1, 2, 3. For most wells (L-1, L-2H, O-2, O-3, K-1, T-1, W-1) the adopted criterion caused the elimination of variables S1, S2, PC and RC from further analysis. For the G-1 and Z-1 wells the correlation allowed the removal of factor variables S2, PC and RC, while for the B-1 well no variable fulfilled the assumed condition of correlation (see Table 4).

Table 1. The matrix of correlation coefficients for measurement variables - the L-1 borehole

Variable	T <sub>max</sub>	S1	S2	S3	PI	PC	RC	TOC	HI	OI	MINC
T <sub>max</sub>	1.00	0.15	0.24	-0.31	-0.32	0.22	0.29	0.28	-0.16	-0.12	-0.40
<b>S</b> 1	0.15	1.00	0.92	-0.16	0.22	0.95	0.86	0.89	0.46	-0.48	-0.23
S2	0.24	0.92	1.00	-0.18	0.00	0.98	0.96	0.97	0.44	-0.43	-0.24
S3	-0.31	-0.16	-0.18	1.00	-0.09	-0.16	-0.17	-0.17	-0.17	0.58	0.44
PI	-0.32	0.22	0.00	-0.09	1.00	0.06	0.00	0.01	0.08	-0.48	0.03
PC	0.22	0.95	0.98	-0.16	0.06	1.00	0.94	0.96	0.44	-0.44	-0.23
RC	0.29	0.86	0.96	-0.17	0.00	0.94	1.00	1.00	0.25	-0.43	-0.26

# NAFTA-GAZ

ect. Table 1

Variable	T <sub>max</sub>	S1	S2	S3	PI	PC	RC	TOC	HI	OI	MINC
TOC	0.28	0.89	0.97	-0.17	0.01	0.96	1.00	1.00	0.29	-0.43	-0.25
HI	-0.16	0.46	0.44	-0.17	0.08	0.44	0.25	0.29	1.00	-0.35	0.07
OI	-0.12	-0.48	-0.43	0.58	-0.48	-0.44	-0.43	-0.43	-0.35	1.00	0.37
MINC	-0.40	-0.23	-0.24	0.44	0.03	-0.23	-0.26	-0.25	0.07	0.37	1.00

Table 2. The matrix of correlation coefficients for measurement variables - the O-2 borehole

Variable	$T_{max}$	S1	S2	S3	PI	PC	RC	TOC	HI	OI	MINC
$T_{max}$	1.00	0.51	0.51	-0.41	0.02	0.52	0.49	0.50	0.08	-0.62	-0.20
<b>S</b> 1	0.51	1.00	0.91	-0.47	0.23	0.94	0.76	0.80	0.19	-0.52	-0.29
S2	0.51	0.91	1.00	-0.39	0.06	0.99	0.92	0.94	0.17	-0.49	-0.29
S3	-0.41	-0.47	-0.39	1.00	-0.55	-0.39	-0.37	-0.38	-0.01	0.68	0.15
PI	0.02	0.23	0.06	-0.55	1.00	0.09	0.02	0.03	-0.14	-0.28	0.29
PC	0.52	0.94	0.99	-0.39	0.09	1.00	0.89	0.92	0.17	-0.49	-0.30
RC	0.49	0.76	0.92	-0.37	0.02	0.89	1.00	1.00	-0.02	-0.47	-0.26
TOC	0.50	0.79	0.94	-0.38	0.03	0.92	1.00	1.00	0.01	-0.48	-0.27
HI	0.08	0.19	0.17	-0.01	-0.14	0.17	-0.02	0.01	1.00	-0.12	-0.21
OI	-0.62	-0.52	-0.49	0.68	-0.28	-0.49	-0.47	-0.48	-0.12	1.00	0.31
MINC	-0.20	-0.29	-0.29	0.15	0.29	-0.30	-0.26	-0.27	-0.21	0.31	1.00

Table 3. The matrix of correlation coefficients for measurement variables - the W-1 borehole

Variable	T <sub>max</sub>	S1	S2	S3	PI	PC	RC	TOC	HI	OI	MINC
$T_{max}$	1.00	0.33	0.21	-0.24	0.03	0.21	0.35	0.34	-0.18	-0.41	-0.06
<b>S</b> 1	0.33	1.00	0.89	-0.29	0.01	0.87	0.94	0.95	-0.10	-0.47	-0.22
<b>S</b> 2	0.21	0.89	1.00	-0.03	-0.35	0.97	0.84	0.85	0.20	-0.18	-0.16
S3	-0.24	-0.29	-0.03	1.00	-0.47	0.09	-0.23	-0.22	0.31	0.67	0.28
PI	0.03	0.01	-0.35	-0.47	1.00	-0.34	-0.02	-0.04	-0.55	-0.48	-0.13
PC	0.21	0.87	0.97	0.09	-0.34	1.00	0.84	0.85	0.16	-0.14	-0.11
RC	0.35	0.94	0.84	-0.23	-0.02	0.84	1.00	1.00	-0.20	-0.47	-0.14
TOC	0.34	0.95	0.85	-0.22	-0.04	0.85	1.00	1.00	-0.19	-0.46	-0.14
HI	-0.18	-0.10	0.20	0.31	-0.55	0.16	-0.20	-0.19	1.00	0.41	0.00
OI	-0.41	-0.47	-0.18	0.67	-0.48	-0.14	-0.47	-0.46	0.41	1.00	0.34
MINC	-0.06	-0.22	-0.16	0.28	-0.13	-0.11	-0.14	-0.14	0.00	0.34	1.00

Table 4. The matrix of correlation coefficients for measurement variables - the B-1 borehole

Variable	T <sub>max</sub>	S1	S2	S3	PI	PC	RC	TOC	HI	OI	MINC
T <sub>max</sub>	1.00	0.41	0.29	-0.42	0.40	0.30	0.40	0.15	-0.29	-0.63	-0.27
S1	0.41	1.00	0.95	-0.32	0.24	0.87	0.98	0.24	-0.12	-0.48	-0.20
S2	0.29	0.95	1.00	-0.26	0.01	0.89	0.95	0.21	0.07	-0.35	-0.18
S3	-0.42	-0.32	-0.26	1.00	-0.17	-0.21	-0.31	-0.07	0.20	0.64	0.53
PI	0.40	0.24	0.01	-0.17	1.00	-0.03	0.19	-0.12	-0.49	-0.38	0.03
PC	0.30	0.87	0.89	-0.21	-0.03	1.00	0.87	0.62	0.01	-0.37	-0.17
RC	0.40	0.98	0.95	-0.31	0.19	0.87	1.00	0.25	-0.16	-0.46	-0.19
TOC	0.15	0.24	0.21	-0.07	-0.12	0.62	0.25	1.00	-0.10	-0.21	-0.07
HI	-0.29	-0.12	0.07	0.20	-0.49	0.01	-0.16	-0.10	1.00	0.28	-0.03
OI	-0.63	-0.48	-0.35	0.64	-0.38	-0.37	-0.46	-0.21	0.28	1.00	0.44
MINC	-0.27	-0.20	-0.18	0.53	0.03	-0.17	-0.19	-0.07	-0.03	0.44	1.00

### **Factor analysis**

Because each measurement variable subject to the analysis is expressed in different units (and can therefore take values from a different range), before conducting the procedure of factor analysis it is necessary to standardise the data. The standardisation was performed in accordance with the following formula:

$$x_i' = \left(\frac{x_i - \overline{x}}{S_x}\right)$$

where:

- $x'_i$  standardised measurement variable,
- x the average value of all the data for the given measurement variable,
- $S_x$  the standard deviation of all the data for the given measurement variable.

As a result of standardisation, we get a space of measurement variables in which the distance between the measurements does not depend on the coordinates. By doing so we eliminate situations in which the given measurement variable has a greater impact on the result of the analysis, because it is expressed as higher absolute values.

The next stage is the replacement of variables remaining after the elimination of correlated data with the so-called factor variables, meaning uncorrelated vectors containing the highest possible amount of information about the variability of input vectors. An important problem here is the choice of the number of factor variables sufficient to credibly describe the variability of measurement data. Taking into account too small a number of variables may cause the loss of information about the variability of the analysed data; on the other hand, taking into account too high a number of them makes the sense of the whole analysis questionable. Table 5 presents a comparison of the number of factor variables obtained for each analysed well for both used criteria.

In most wells the numbers of factor variables for the Kaiser and Cattel criterions are in mutual compliance (the B-1, L-2H, K-1, O-2 and O-3 wells) or do not differ considerably (W-1, Z-1). The G-1 and T-1 wells, for which the difference in the number of variables depending on the adopted criterion amounts to two, are the exception. The plot with the eigenvalues for these wells are presented in Figures 1 and 2.

In accordance with the Kaiser criterion, for both wells we get two factor variables (eigenvalues higher than 1.0), with four for the Cattel criterion, it being possible for the T-1 well to adopt even five factor variables depending on the method of the interpretation of the plot. A separate case is the L-1 well,

Borehole	The Kaiser criterion	The Cattel criterion
B-1	3	3
L-1	2/3	does not resolve
L-2H	2	2
G-1	2	4
K-1	2	2
O-2	2	2
O-3	2	2
T-1	2	4/5
W-1	2	3
Z-1	3	4

Table 5. The number of factor variables -

the comparison of criteria



for which it is not possible to resolve the number of factor variables according to the Cattel criterion (Fig. 3).

It seems that in this situation adopting the percentage criterion would be the most effective, according to which we take

# NAFTA-GAZ





into account as many factor variables as possible to explain the arbitrarily selected percentage of the variance of original variables. Table 6 presents the percentage of variability taken into account by two, three and four factor variables and the eigenvalues obtained during the analysis.

Adopting the assumption that the same number of factor variables should be chosen for each well, it would be the most effective to choose three variables, which, on the one hand, means a considerable reduction of the number of original variables, and on the other hand, it causes the resulting fac-

Doroholo	Figenvalues	The number of factor variables						
Borenoie	Eigenvalues	2	3	4				
B-1	4.63 / 2.04 / 1.04	66.73%	77.13%	86.97%				
L-1	2.18 / 1.48 / 0.99	61.03%	77.69%	90.19%				
L-2H	2.58 / 1.23 / 0.87	63.49%	78.05%	89.88%				
G-1	3.02 / 2.18 / 0.95	74.22%	87.79%	93.11%				
K-1	3.56 / 1.00 / 0.70	76.04%	87.75%	94.45%				
O-2	2.54 / 1.20 / 0.85	62.25%	76.48%	87.85%				
O-3	2.67 / 1.28 / 0.95	65.80%	81.57%	90.26%				
T-1	2.61 / 1.74 / 0.97	72.60%	88.69%	95.85%				
W-1	2.73 / 1.29 / 0.73	67.05%	79.14%	90.20%				
Z-1	2.55 / 1.90 / 1.25	63.68%	81.56%	89.57%				

Table 6. The number of factor variables and the eigenvalues

tor variables to take into account the majority of information included in the original data. Although for most wells the use of the Kaiser and Cattel criteria indicated taking into account two factor variables, the analysis of the variability of vectors taken under consideration (Tab. 6) decided about the selection of three variables. Ultimately, the conducted analysis allowed an over threefold reduction of the number of measurement variables, at the same time retaining over 75% of information about the variability of parameters for each well.

### **Summary**

The paper verifies the possibility of using factor analysis for the purpose of the credible reduction of the amount of measurement variables, based on the example of physicochemical data from the wells in shale formations. The conducted statistical analysis indicated the effectiveness of the used method, and the results obtained during the analysis allow the formulation of the following conclusions:

- The factor analysis is a method of data analysis allowing the effective reduction of the number of measurement variables obtained during various kinds of research, and the physicochemical data constitute a coherent set of information making it possible to conduct a complete statistical analysis.
- The conducted two-stage reduction of the number of measurement variables allowed an over threefold reduction of their number of measurement variables while retaining over 75% of information about their variability.
- The procedure of determining the number of factor vari-

ables should not be based on one of the criteria which are possible to be used, but on all the information acquired as a result of the analysis, i.e. the eigenvalues, the percentage of the explained variance of data, the monotonicity of the "scree test".

- Due to the diversity of the properties of the rocks depending on the time and the mechanism of their generation, the result of a factor analysis may considerably depend on the selection of the stratigraphic unit for which the analysis is being conducted, which means the necessity to conduct it independently for each stratigraphic unit represented in the well.
- The values of factor variables for each well constitute a point of reference for further analyses, in particular for the use of agglomeration analysis for the purpose of categorisation, meaning the grouping of wells with respect to their statistical similarity.

Please cite as: Nafta-Gaz 2016, no. 12, pp. 1069-1075, DOI: 10.18668/NG.2016.12.09

Article contributed to the Editor 1.04.2016. Approved for publication 20.07.2016.

The paper is the result of research conducted as part of the project: *Methodology of determining sweet spots on the basis of geochemical, petrophysical and geomechanical properties, based on the correlation of the results of laboratory examinations with the* 

# artykuły

geophysical measurements and the 3D generational model. Subject 8: the construction of a computer system for storing and processing information involving shale formations as the potential deposits of unconventional hydrocarbons, co-funded by the National Centre for Research and Development as part of the Blue Gas – Polish Shale Gas programme. Contract No. BG1/MWSSSG/13.

## Literature

- [1] Budak P., Łętkowski P., Szpunar T., Nowak R., Radzikowski K., Arabas J.: SweetSpot – a computer system for storing and sharing data on the rock properties of shale formations. Nafta-Gaz 2015, no. 12, pp. 944–952, DOI: 10.18668/NG2015.12.02.
- [2] Filar B., Kwilosz T., Miziołek M., Piesik-Buś W., Zamojcin J.: The use of cluster analysis for the segmentation of the physi-

*cochemical properties of shale gas deposits*. Nafta-Gaz 2015, no. 11, pp. 898–909, DOI: 10.18668/NG2015.11.13.

[3] Klaja J., Łykowska, G.: Wyznaczenie typów petrofizycznych skał czerwonego spągowca z rejonu południowo-zachodniej części niecki poznańskiej na podstawie analizy statystycznej wyników pomiarów laboratoryjnych. Nafta-Gaz 2014, no. 11, pp. 757–764.



Dr. Eng. Piotr ŁĘTKOWSKI PhD. Assistant Professor Department of Hydrocarbon Deposits and UGS Facilities Simulation Oil and Gas Institute – National Research Institute ul. Lubicz 25 A, 31-503 Kraków E-mail: *piotr.letkowski@inig.pl* 



Paweł BUDAK M.Sc. Eng. Senior Science and Research Specialist Head of the Department of Petroleum Engineering Oil and Gas Institute – National Research Institute ul. Lubicz 25 A 31-503 Kraków E-mail: *pawel.budak@inig.pl* 



Andrzej GOŁĄBEK M.Sc. Eng. Junior Scientist Department of Hydrocarbon Deposits and UGS Facilities Simulation Oil and Gas Institute – National Research Institute ul. Lubicz 25 A, 31-503 Kraków E-mail: *andrzej.golabek@inig.pl* 



Associate Professor

Prof. dr hab. Eng. Jarosław ARABAS

ul. Nowowiejska 15/19, 00-665 Warszawa

Department of Electronics and Information Technology

Warsaw University of Technology

Institute of Electronic Systems

E-mail: jarabas@ise.pw.edu.pl

Dr. Eng. Tadeusz SZPUNAR PhD. Assistant Professor Department of Petroleum Engineering Oil and Gas Institute – National Research Institute ul. Lubicz 25 A 31-503 Kraków E-mail: *tadeusz.szpunar@inig.pl* 



Dr hab. Eng. Robert NOWAK Lecturer/Assistant Professor Warsaw University of Technology Department of Electronics and Information Technology Institute of Electronic Systems ul. Nowowiejska 15/19, 000-665 Warszawa E-mail: *R.M.Nowak@elka.pw.edu.pl*