Multidiszciplináris tudományok, 12. kötet. (2022) 3 sz. pp. 89-106 **https://doi.org/10.35925/j.multi.2022.3.9**

CORRELATION-BASED IMPUTATION METHOD FOR ESTIMATING MISSING WELL LOG DATA IN A HUNGARIAN GROUNDWATER WELL

Viktória Kiss
PhD student, University of Miskolc, Institute of Geophysics and Geoinformatics, Department of Geophysics
3515 Miskolc-Egyetemváros, e-mail: kviktoria246@gmail.com

Norbert Péter Szabó
Full professor, University of Miskolc, Institute of Geophysics and Geoinformatics, Department of Geophysics
3515 Miskolc-Egyetemváros, e-mail: gfnmail@uni-miskolc.hu

Abstract

Incomplete well logging datasets measured in boreholes are frequently encountered especially in old wells. The main goal of the research, described in this paper, is to refill a data matrix with synthetic data that is missing for some reason and how the imputation procedure was to determine well the measured and estimated values fit in the case of data gaps of different sizes, and how the clay content and porosity change as a result. There can be many reasons for the lack of data, such as not measuring the desired parameters in a given depth zone. To fill this data gap, correlation-based imputation method was used which was applied in MATLAB software development system. A case study involving a Hungarian thermal water well is shown to demonstrate the reliability of the multi-linear correlation based method, which can be fruitfully applied in other wells and investigation areas.

Keywords: *missing value, data matrix, correlation, imputation, well log*

1. Introduction

In geophysical measurements, there is almost never a case where no data is missing from the measured dataset. In case of different data deficiencies, the application of the imputation method is very efficient, as it can make up for the missing values. This method is based on the incomplete data matrix they are replenished by estimation (Rubin, 1987), thus, statistical analyzes can be performed. To understand the structure of missing data, we divide the data matrix into two parts: the known part and the missing part, Y = (Yobserved, Ymissing) (Rubin, 1976). In this study, an imputation procedure is presented, that is a very common problem in various researches. Estimating missing data is of great importance in statistical analysis because an incomplete data set does not give the correct answer, and deficiencies degrade the quality and reliability of the analysis (Máder, 2005).

One such efficient imputation procedure is based on correlation calculations, which was suggested by Szabó et al. (2019) for completing big matrices of core measurement data. We implement the same technique for estimating missing wireline logging data. A more detailed mathematical explanation can be found in the aformentioned article. In this case, we explain the application of the imputation method using well logging data contained in an incomplete data matrix. In the first step, by using an own developed, MATLAB software, value gaps can be filled with the so-called NaN (Not-a-Number) values. The point of the method is to select the two data columns that correlate best with each other. This is

possible if there are data sets in which we find values in both rows, thus the method estimates the next data pair based on them.

Table 1 shows an example where some values of a spotaneous potential and a natural gamma-ray intensity data are missing. We marked the lines where one data is known and the other is missing in yellow, thus in this case the imputation procedure can be performed. The new values are estimated using a linear regression relationship between the relevant variables established on the existing data. In the red line, it is not possible to replace the missing values. A new column should be involved for the calculations. The imputation procedure is performed on the entire data matrix, examining all possible pairs of measured variables. Data replacement in the given row is always carried out using the two most strongly correlated measurement variables (columns *i* and *j*), assuming a linear relationship between the variables. (Linearity criterion is acceptable in a small domain of the independent variable.) *Equation (1)* contains the regression model, which gives the estimated regression coefficients (β) ($j \neq i$)

$$Y_j = \beta_0 + \beta_1 Y_i. \tag{1}$$

We apply the imputation algorithm to the input data matrix until it is completely filled with the values extracted from the linear equations (Dong et al., 2013; Rubin, 1987).

Table 1. *An example of the scheme of the suggested imputation procedure in case of different data gaps*

Depth [m]	Spontaneous potential [mV]	Natural gamma-ray intensity [API]
413	–32.4	Missing value
414	Missing value	427.8
415	Missing value	Missing value
416	–33.7	438.4

2. Application of correlation-based imputation method to field data

The input of the procedure was a dataset measured in Baktalórántháza-1 well. It included Depth [m], the E40 drill bit, Spotnaeous potential [mV], Natural gamma-ray intensity [GAPI], Gamma-gamma (Density) [g/cm^3], Neutron porosity [%], and Caliper [in] values. We selected an interval from 400 meters to 486.4 meters, this gives a total of 865 depth points.

In the area of Baktalórántháza, there is a Paleogene-Upper Cretaceous sedimentary substrate (Völgyi, 1984). Here there are sandy layers with different lime content with varying degrees of periodic sand movements resulting alluvial cones. These sandy formations contain layers with iron and clay (Buró, 2015).

As we mentioned, we had to select two data columns that correlate best with each other, thus we chose Spontaneous potential (SP) – Natural Gamma-ray intensity (GR), Gamma-gamma intensity (GG) – Neutron-neutron intensity (NN) and Borehole caliper (CAL) – Spontaneous potential pairs. We overloaded the data system with missing values of 10% at first, then 40%. Although, it raises problems above 15% for data gaps (McDermit et al., 1999), we found the above values appropriate for illustration.

2.1. Case 1. – 10% missing data

Implementation of the imputation method was performed using MATLAB software system. On the left side of *Figure 1* we can see the measured data matrix, in the faint places with the missing value, while on the right side we can see the data matrix filled by the correlation procedure, for the seven sections

depending on the depth. The 10% incompleteness is still very small, thus the imputation procedure is very expedient here. During programming we used a pairwise deletion, thus only those cases were currently omitted where is a lack of data according to the variables included (Cool, 2000). We chose 4 as the moving average. In *Figure 2* one can see the measured (green) and estimated (red) sections in terms of depth. The SP and CAL profiles show the best match, while the measured and calculated values of the GR, GG, and NN profiles fit less well to each other.

Figure 3 also shows the measured and the calculated sections, supplemented with the incomplete section. In this case, too, it can be clearly seen that the 10% deficit does not significantly degrade the image of the section. The measured and calculated values fit pretty well, there are only one or two outliers. In *Figure 4* we see the comparison of different sections in relation to each other. In this case, the values are grouped quite well. After plotting, we performed a data correlation shown in *Table 2* (E40, SP, GR, GG, NN, CAL) with an error rate of 21.15%. This error value results of the estimation accuracy of the method, thus the estimated values are substituted for the missing values (Szabó et al, 2019). After all of this, we examined how clay content and porosity changed with different degrees of data gaps. We also created clay content data with 5%, 15% and 25% missing values. The clay content of missing (top) and complete (bottom) dataset are shown in *Figure 5.*

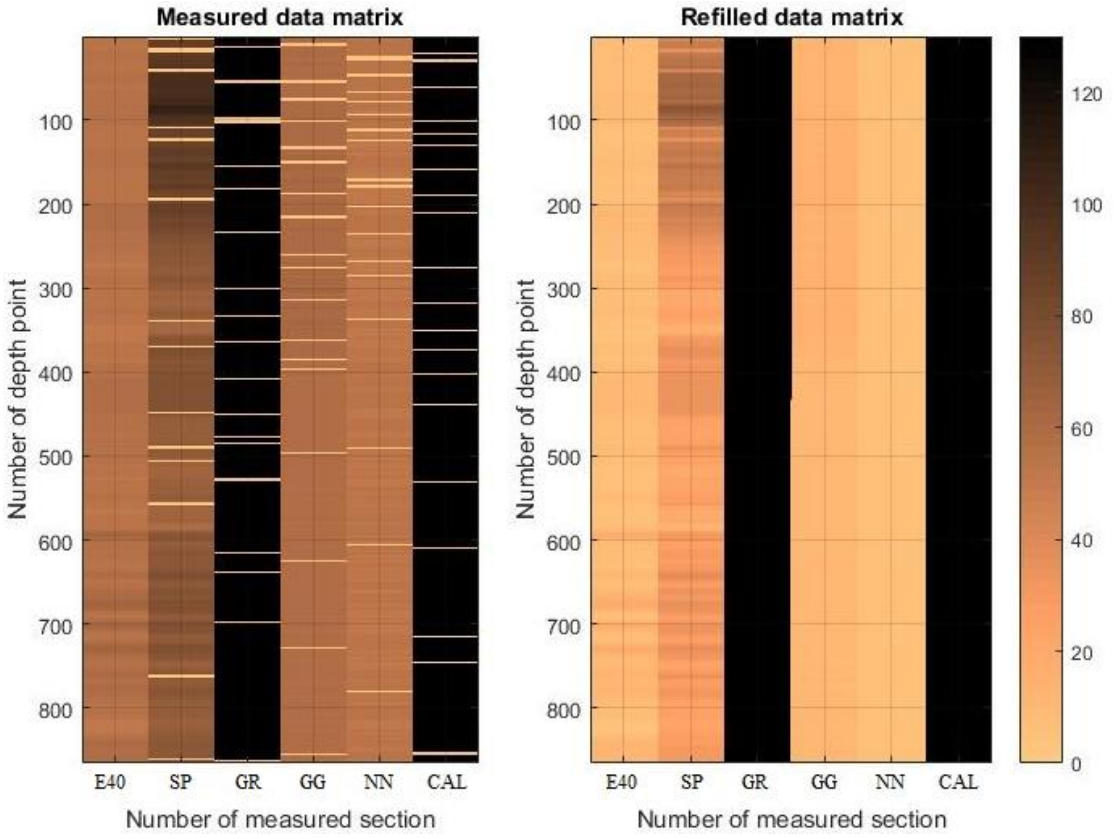


Figure 1. *Measured and refilled data matrix of 10% missing values [the scale on the right shows the degree of correlation: in the sections where the lack of data is large, it is weak (0), where the lack of the data is smaller, it is strong (120)]*

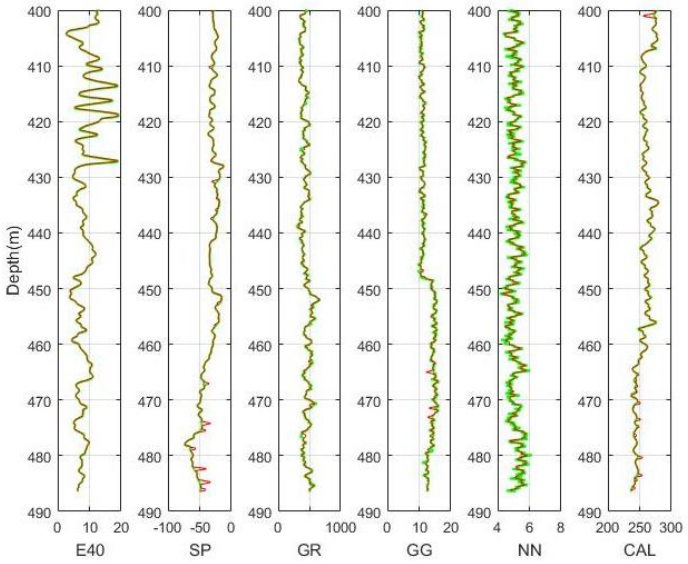
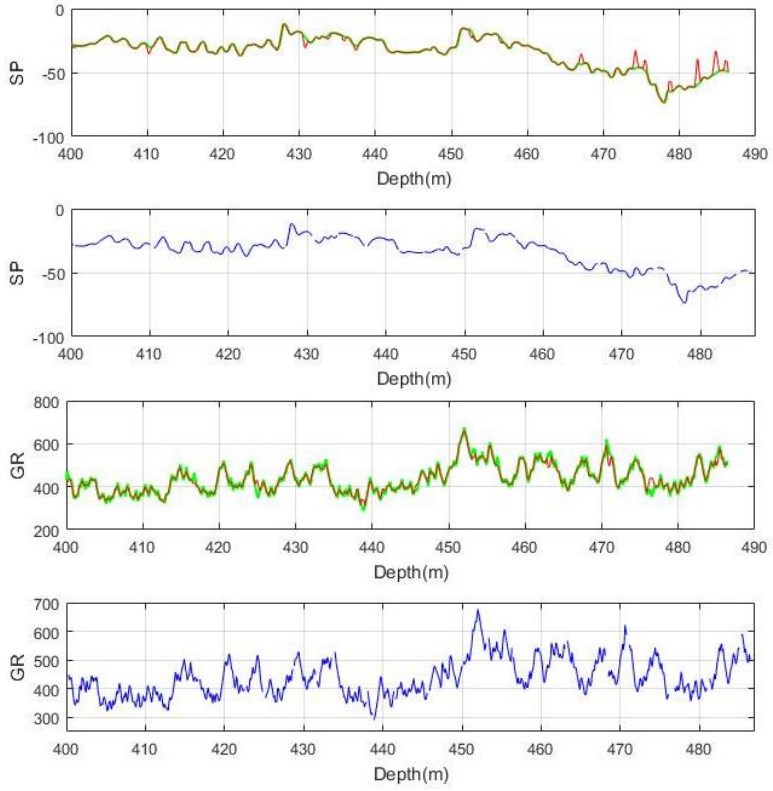


Figure 2. *Measured (green) and estimated (red) sections in terms of depth (from left to right: E40, SP, GR, GG, NN, CAL) for 10% missing values*



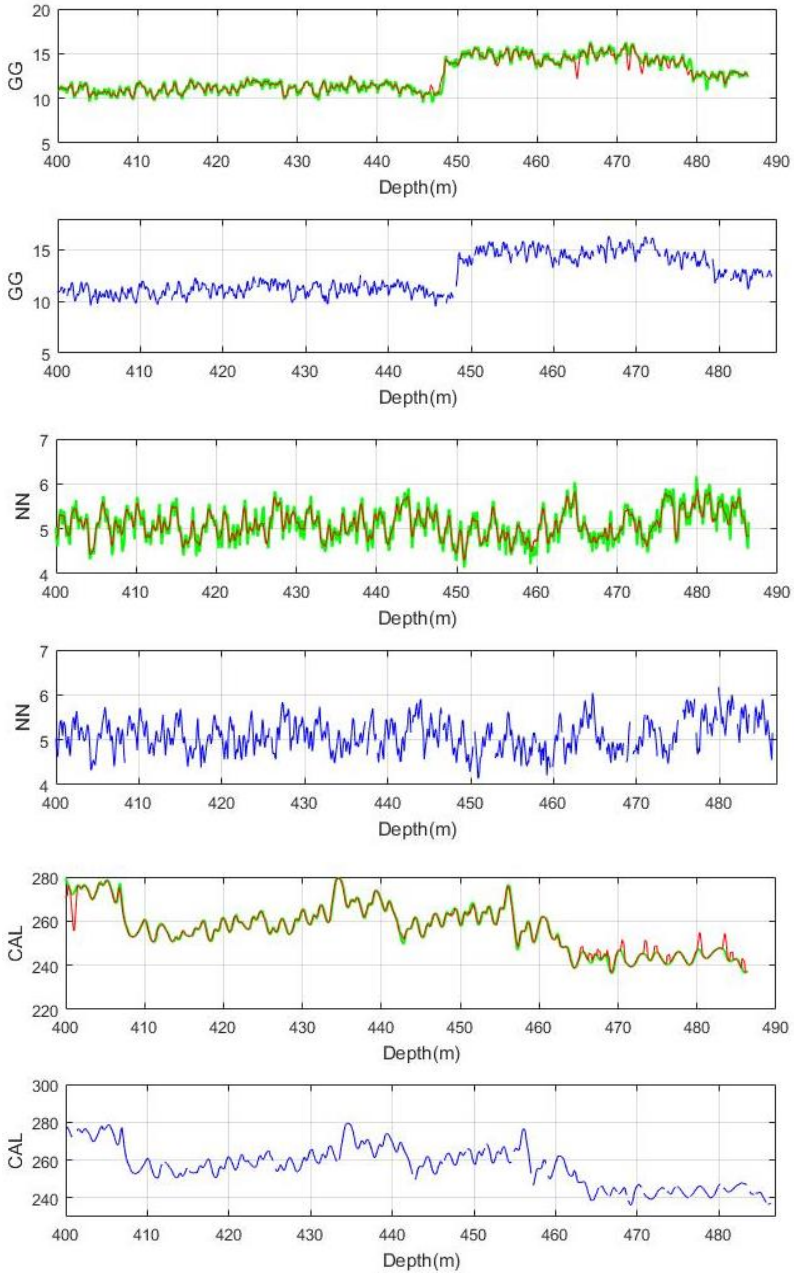


Figure 3. *Measured (green), calculated (red) and incomplete (blue) sections in terms of depth (from top to bottom: E40, SP, GR, GG, NN, CAL) for 10% missing values*

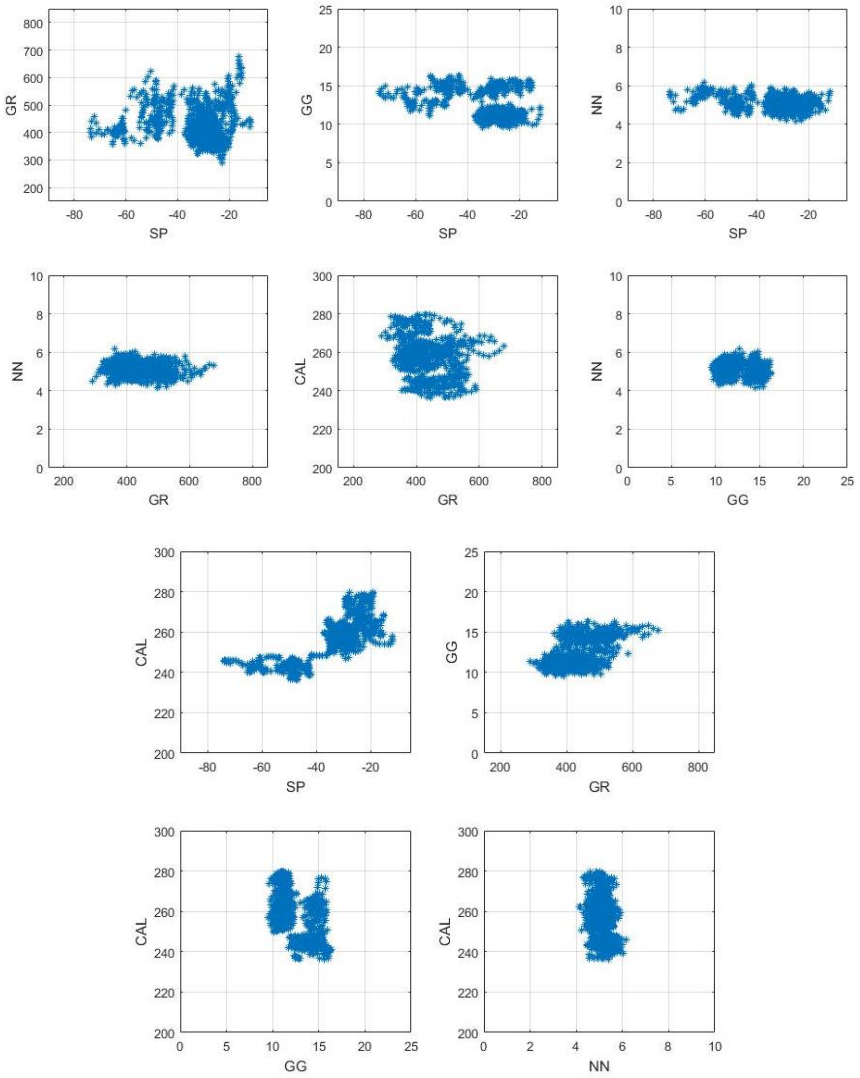


Figure 4. *Comparison of different sections in relation to each other for 10% missing values*

Table 2. *Data correlation matrix for 10% deficit*

1.0000	–0.1068	–0.4086	–0.3190	0.2294	–0.1006
-0.1068	1.0000	–0.0188	–0.3683	–0.3251	0.7511
-0.4086	–0.0188	1.0000	0.5755	–0.1097	–0.2118
-0.3190	–0.3683	0.5755	1.0000	–0.1343	–0.4416
0.2294	–0.3251	–0.1097	–0.1343	1.0000	–0.2576
-0.1006	0.7511	–0.2118	–0.4416	–0.2576	1.0000

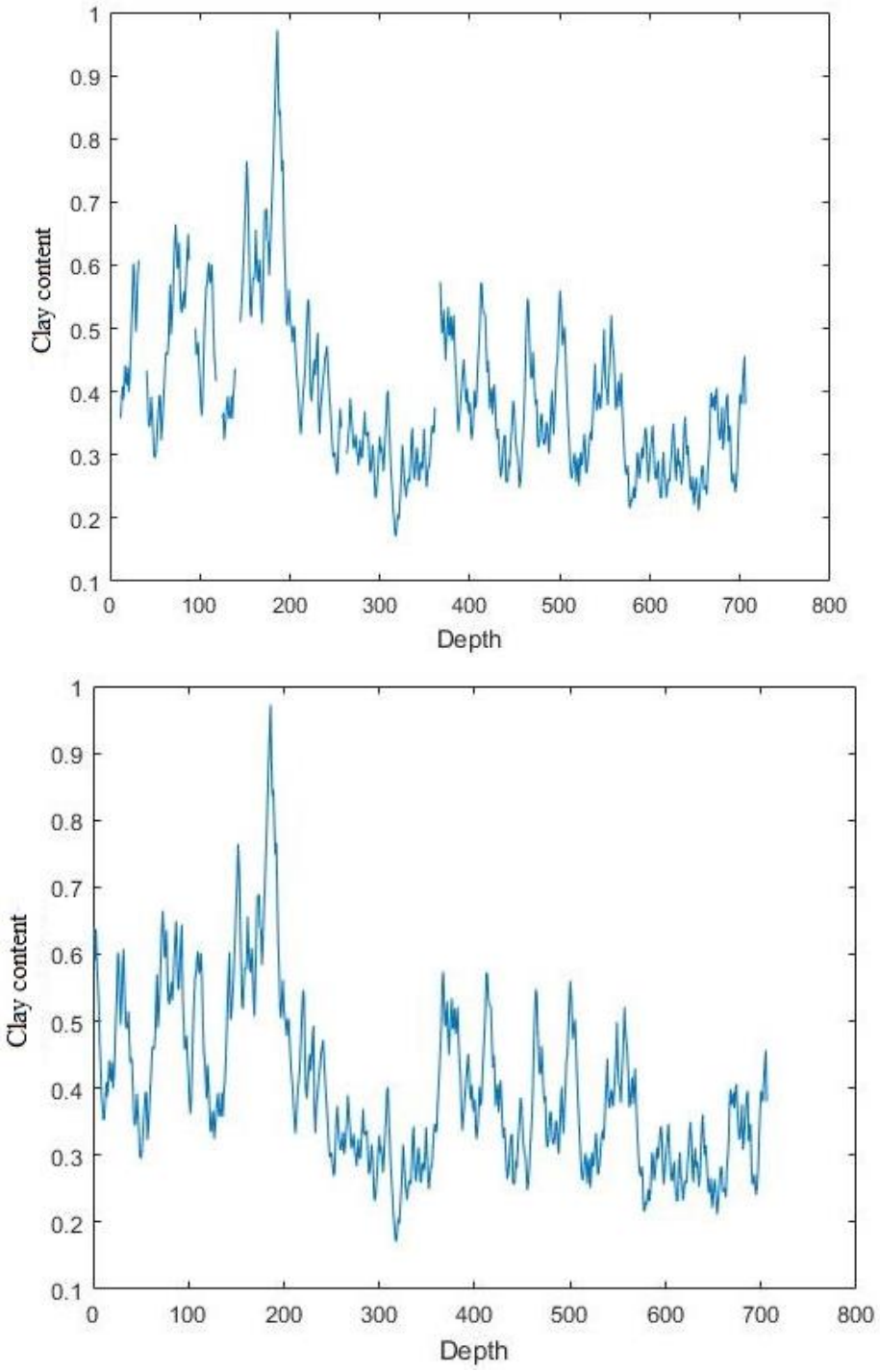


Figure 5. *The clay content of missing (top) and complete (bottom) dataset in terms of depth for 5% missing data*

2.2. Case 2. – 20% missing data

In the second case, we created the dataset with a deficit of 20%. Represented as above: in *Figure 6* we can see the measured and filled data matrix in terms of depth. The 20% deficit cause a little bit higher uncertainty than the 10% but the fit is not yet significantly different. In *Figure 7,* we see the measured and estimated sections in terms of depth. In this case, as well, the SP and CAL profiles show the best match. The other profiles fit less well to each other. *Figure 8* shows the measured and the calculated sections with the incomplete section. The 20% deficit does not greatly distort the profiles. In *Figure 9* we see different sections in relation to each other. We also performed a data correlation shown in *Table 3* (E40, SP, GR, GG, NN, CAL) with an error rate of 23.1%. We showed how the clay content react to missing values (top) and the complete dataset (bottom) in *Figure 10.*

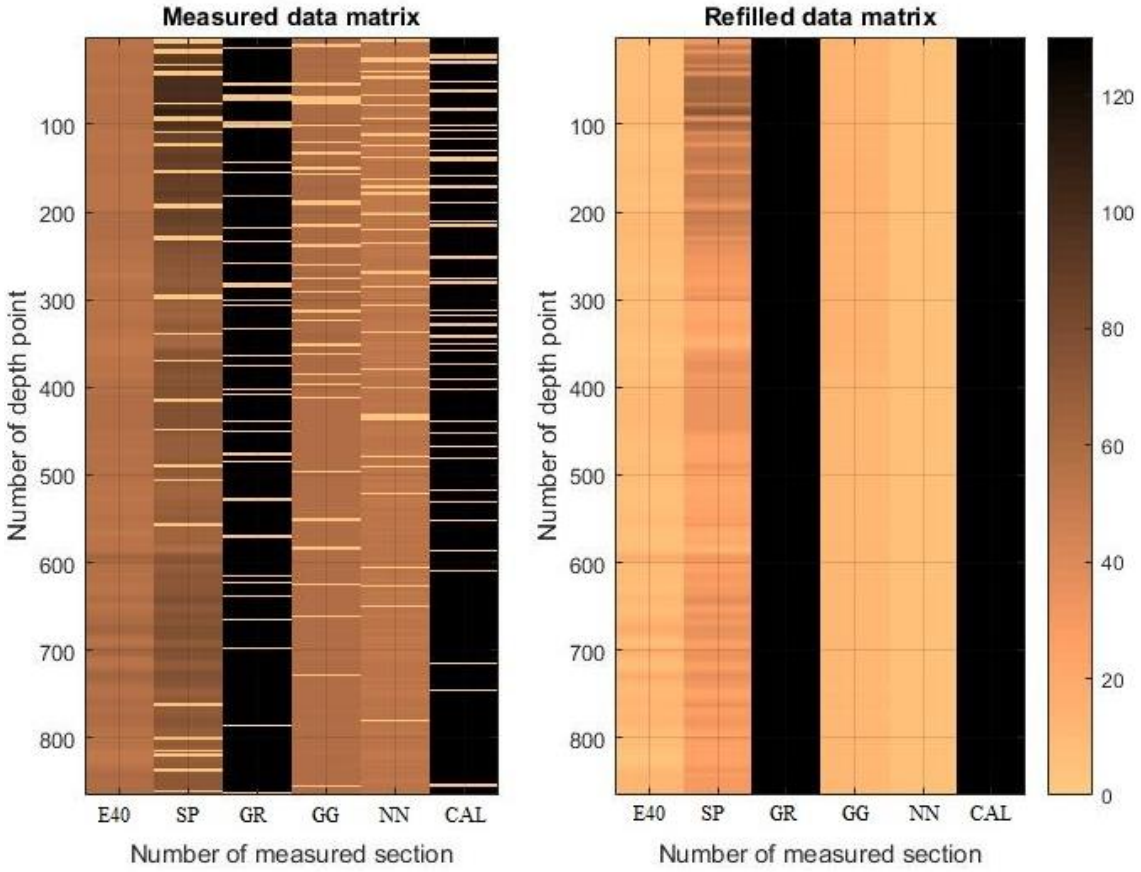


Figure 6. *Measured and refilled data matrix of 20% missing values [the scale on the right shows the degree of correlation: in the sections where the lack of data is large, it is weak (0), where the lack of the data is smaller, it is strong (120)]*

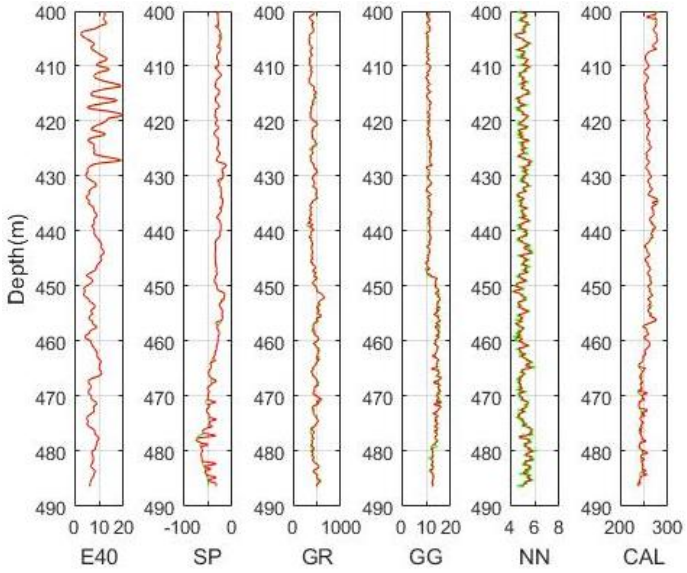
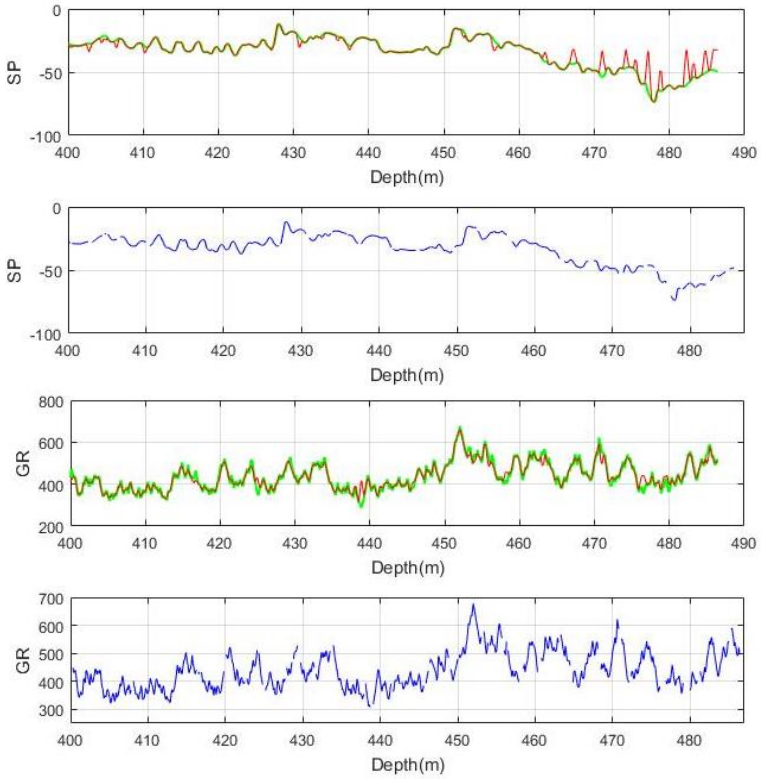


Figure 7. *Measured (green) and estimated (red) sections in terms of depth (from left to right: E40, SP, GR, GG, NN, CAL) for 20% missing data*



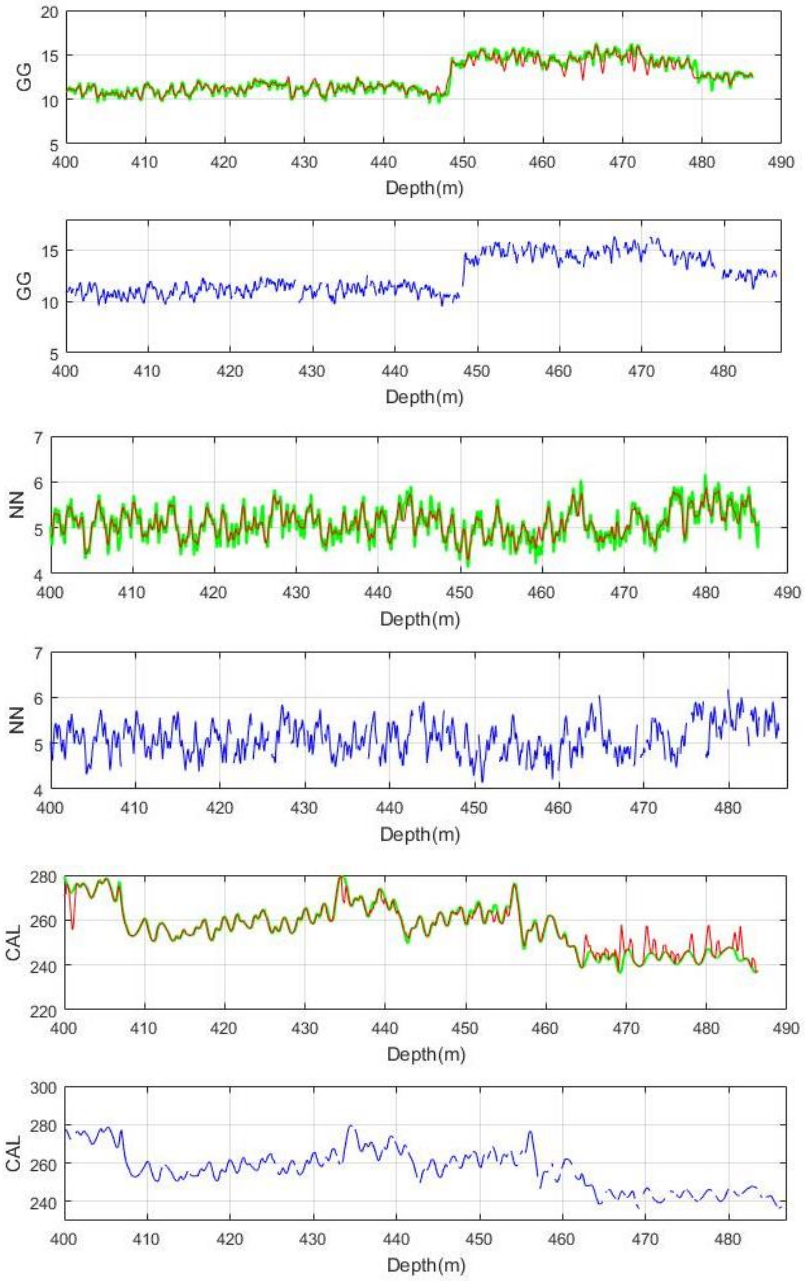


Figure 8. *Measured (green), calculated (red) and incomplete (blue) sections in terms of depth (from top to bottom: E40, SP, GR, GG, NN, CAL) for 20% missing data*

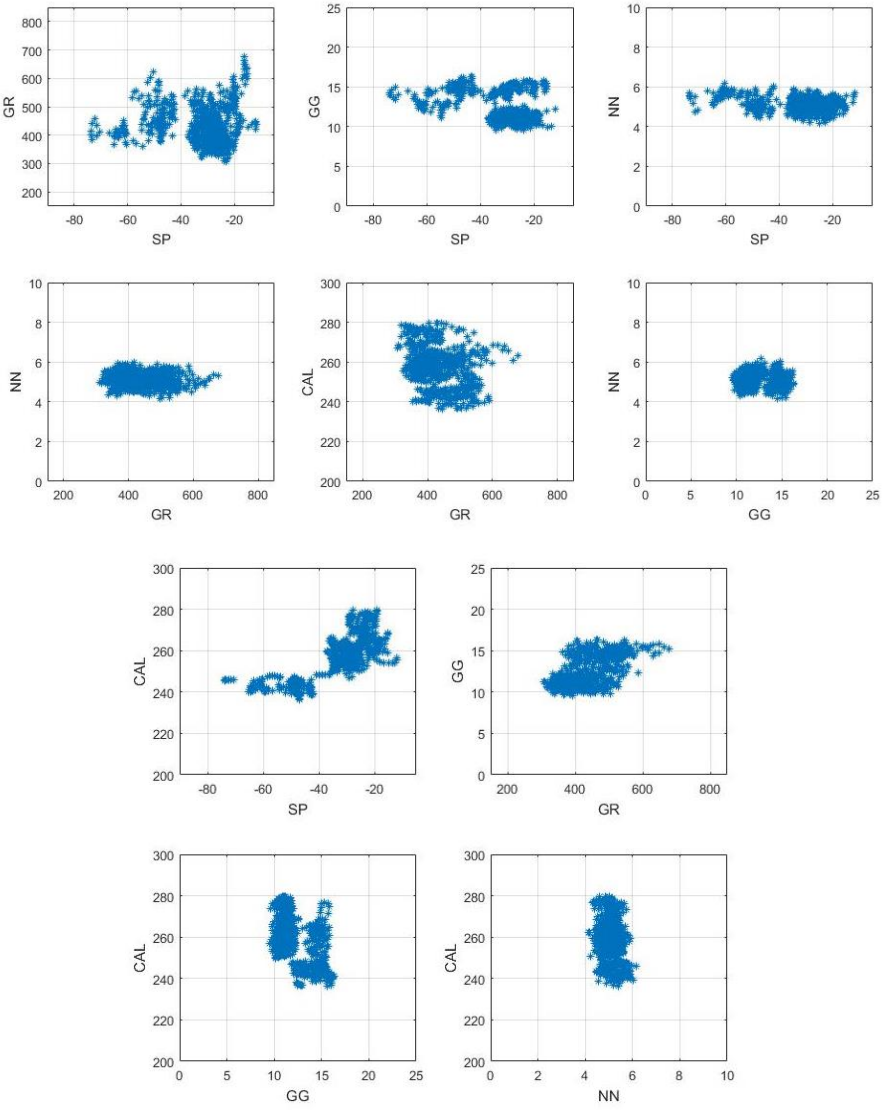


Figure 9. *Comparison of different sections in relation to each other for 20% missing data*

Table 3. *Data correlation matrix for 20% deficit*

1.0000	–0.1909	–0.4192	–0.3176	0.2303	–0.0881
–0.1909	1.0000	–0.0311	–0.2816	–0.3225	0.5289
–0.4192	–0.0311	1.0000	0.5818	–0.0332	–0.3260
–0.3176	–0.2816	0.5818	1.0000	–0.1306	–0.3783
0.2303	–0.3225	–0.0332	–0.1306	1.0000	–0.2116
–0.0881	0.5289	–0.3260	–0.3783	–0.2116	1.0000

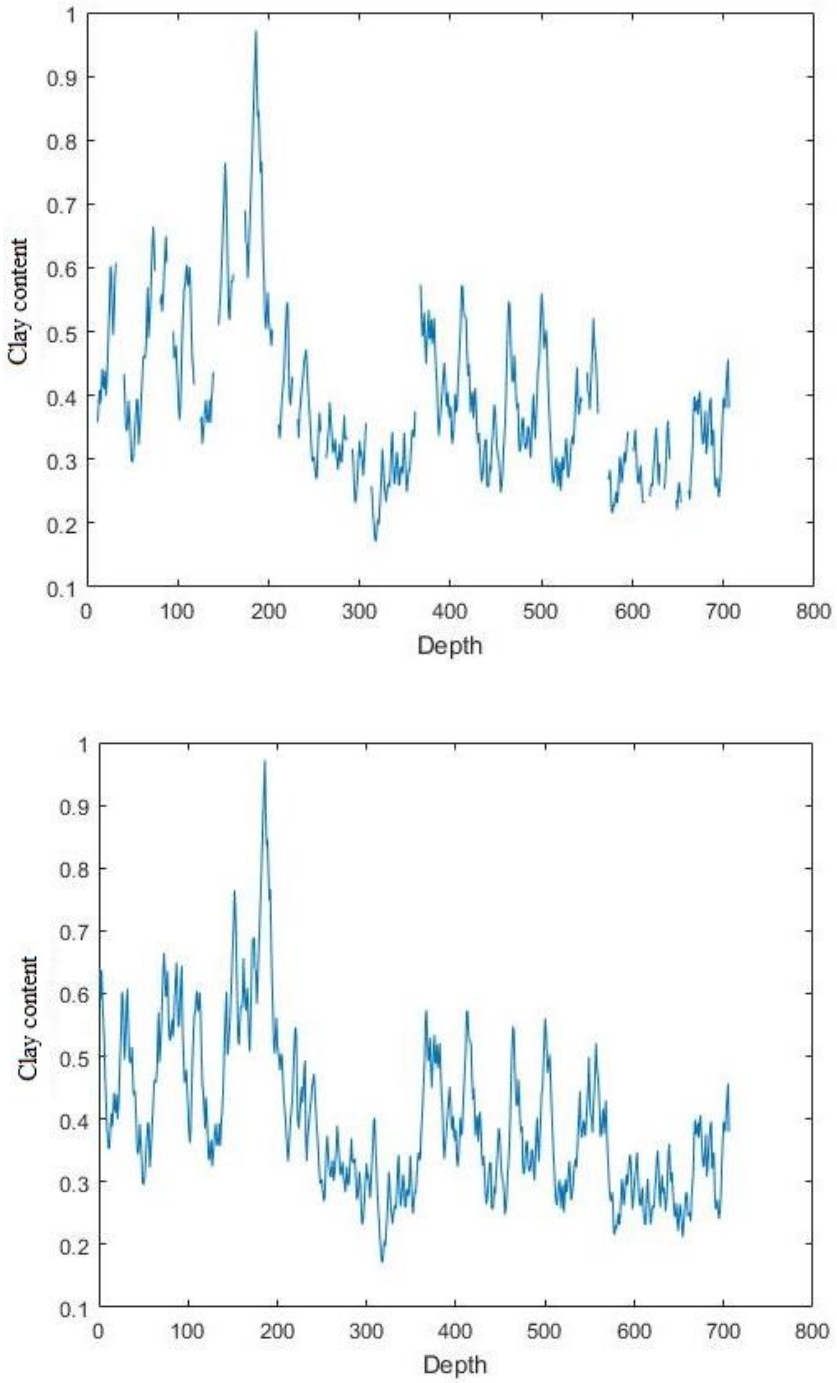


Figure 10. *The clay content of missing (top) and complete (bottom) dataset in terms of depth for 15% missing data*

2.3. Case 3. – 40% missing data

In the third case, the data deficit was 40%. In *Figure 11*, we can see the measured and filled data matrix in terms of depth. The 40% deficit cause a much higher uncertainty in the interpretation. In *Figure 12* we see the measured and estimated sections. In this case, all profiles show much less fit in measured and estimated data. *Figure 13* shows the measured and the calculated sections with the incomplete section. The 40% deficit has higher distortion ability. In this case, there are more outliers, thus the fit also deteriorates. In *Figure 14*, we see different sections in relation to each other, here, the values tend to be scattered, only compressed in one or two places.

The data correlation shown in *Table 4* (E40, SP, GR, GG, NN, CAL) with an error rate of 31.1%. The clay content and the porosity has changed much more compared to the previous ones. The change in missing values of clay content (top) and complete clay content (bottom) is shown in *Figure 15*. It can be seen that the higher the missing value for the clay content, the less it remains good.

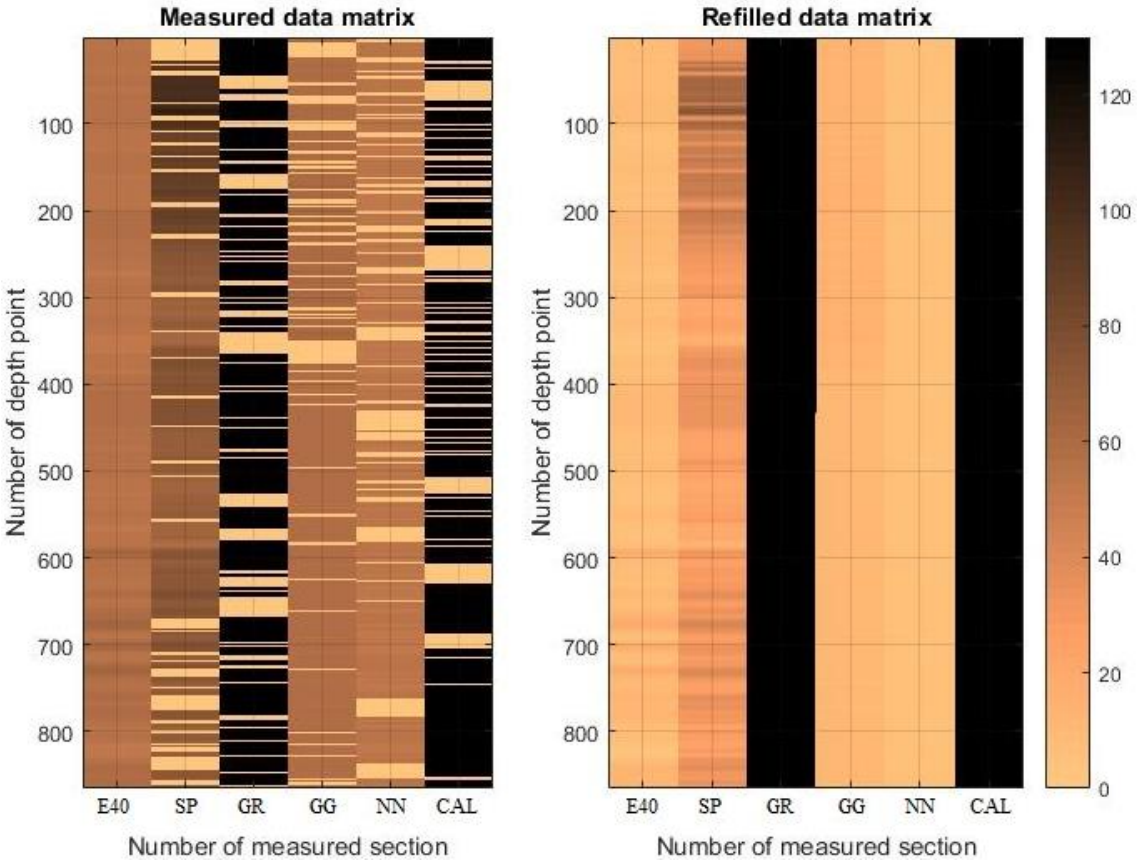


Figure 11. *Measured and refilled data matrix of 40% missing values [the scale on the right shows the degree of correlation: in the sections where the lack of data is large, it is weak (0), where the lack of the data is smaller, it is strong (120)]*

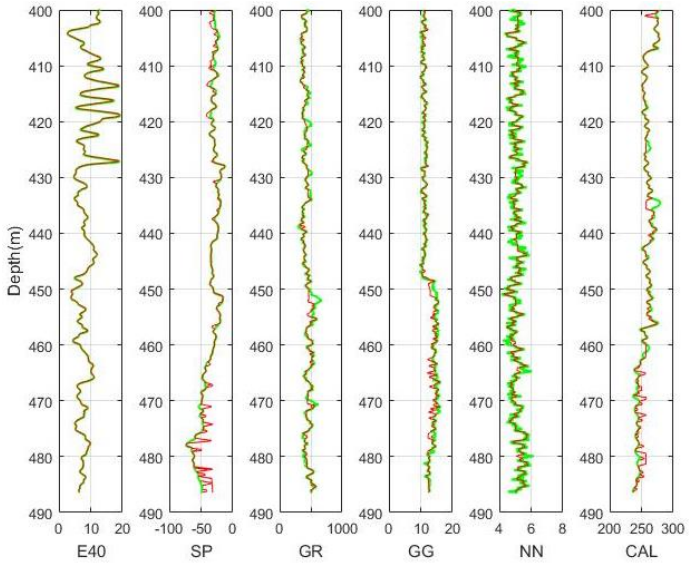
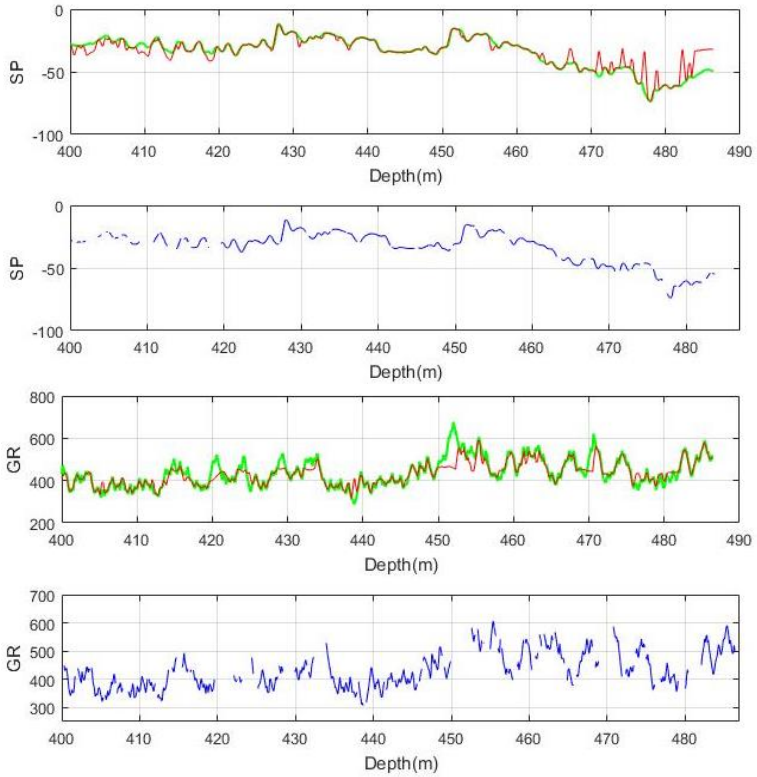


Figure 12. *Measured (green) and estimated (red) sections in terms of depth (from left to right: E40, SP, GR, GG, NN, CAL) for 40% missing values*



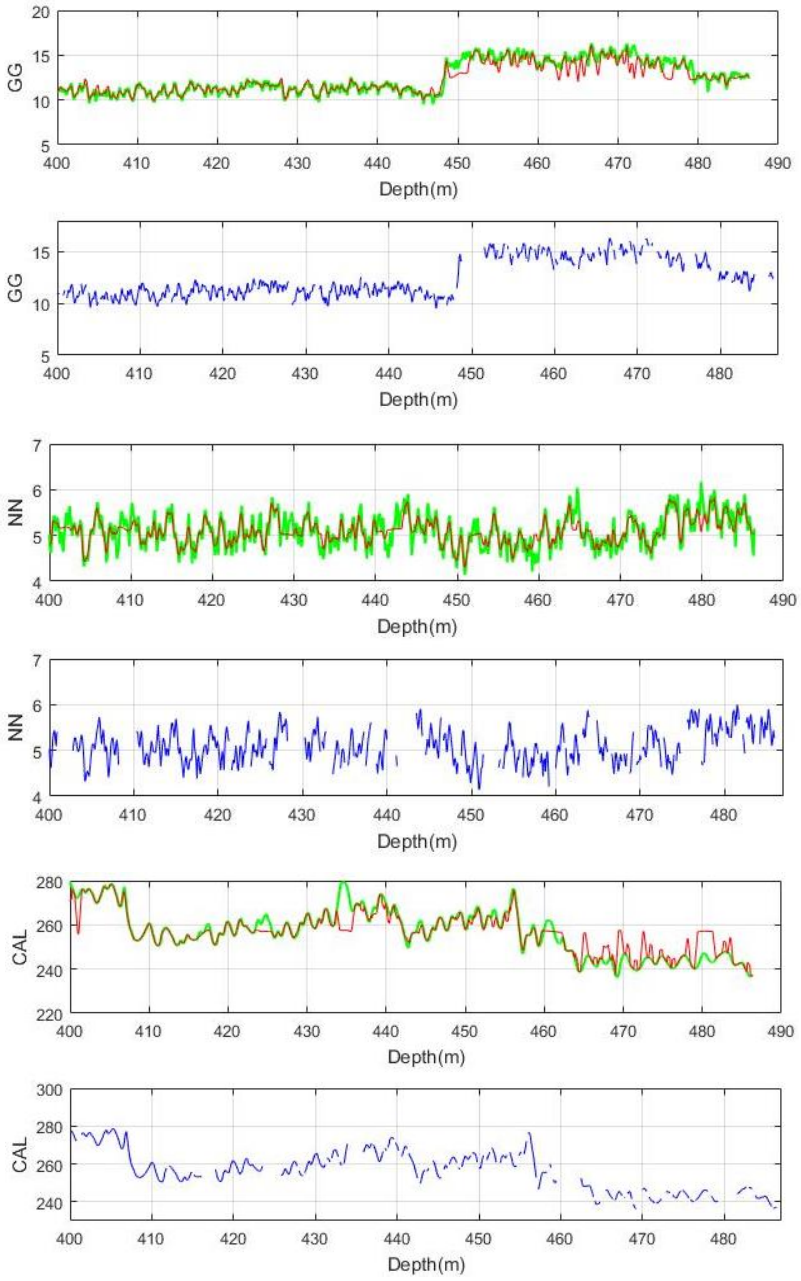


Figure 13. *Measured (green), calculated (red) and incomplete (blue) sections in terms of depth (from top to bottom: E40, SP, GR, GG, NN, CAL) for 40% missing values*

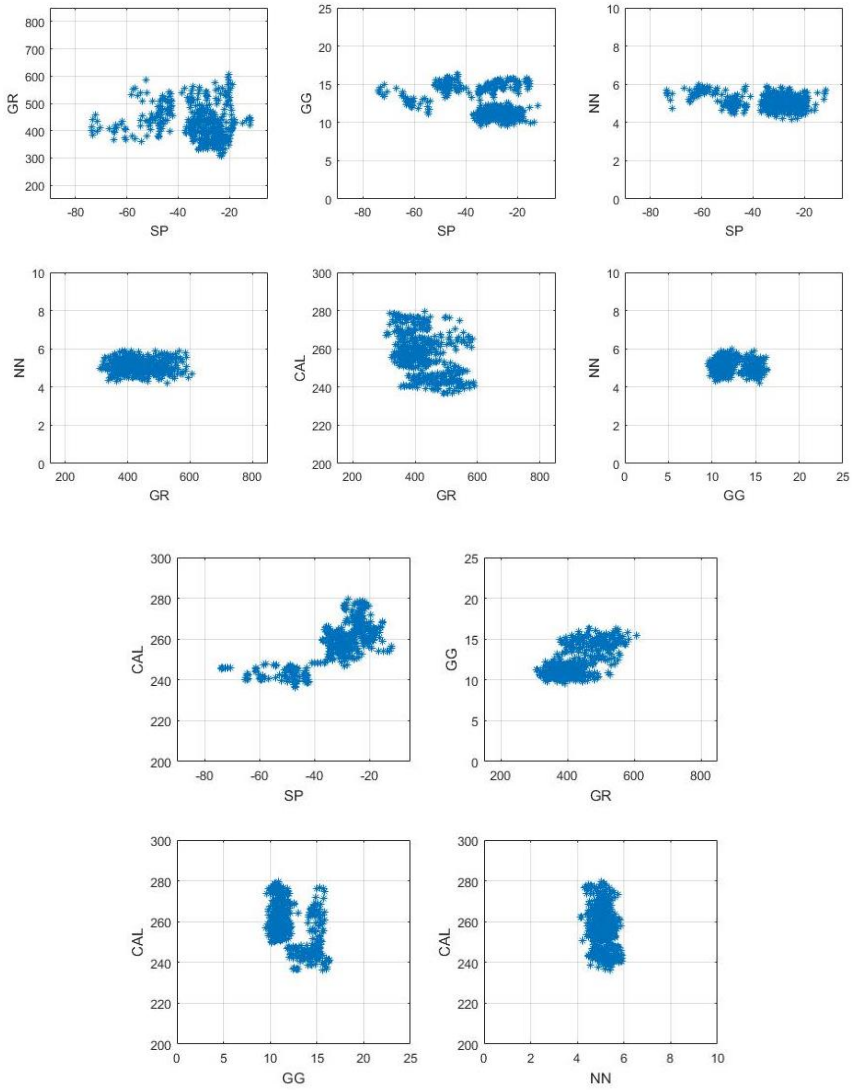


Figure 14. *Comparison of different sections in relation to each other – 40%*

Table 4. *Data correlation matrix for 40% deficit*

1.0000	–0.1909	–0.4192	–0.3176	0.2303	–0.0881
–0.1909	1.0000	–0.0311	–0.2816	–0.3225	0.5289
–0.4192	–0.0311	1.0000	0.5818	–0.0332	–0.3260
–0.3176	–0.2816	0.5818	1.0000	–0.1306	–0.3783
0.2303	–0.3225	–0.0332	–0.1306	1.0000	–0.2116
–0.0881	0.5289	–0.3260	–0.3783	–0.2116	1.0000

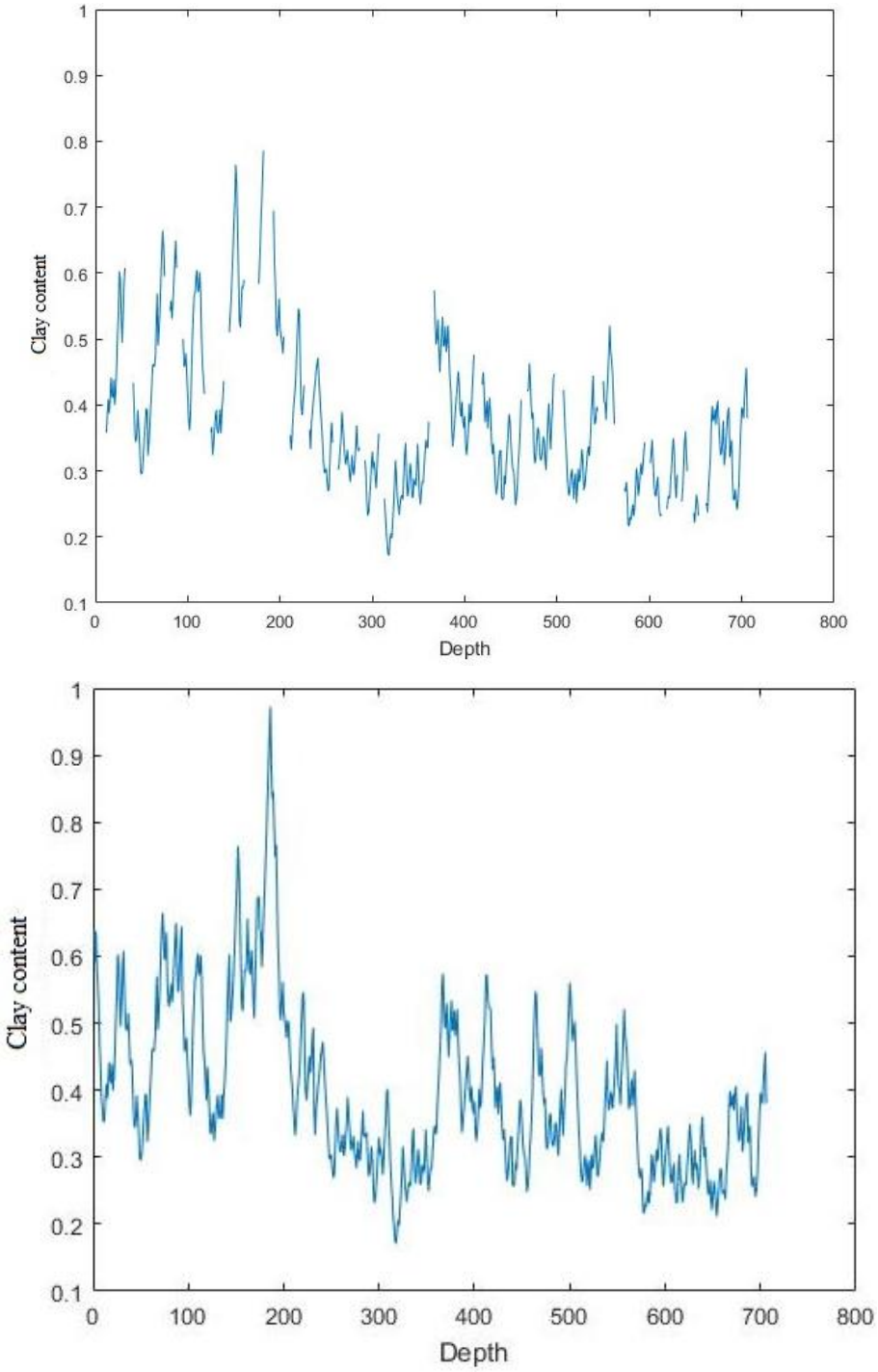


Figure 15. *The clay content of missing (top) and complete (bottom) dataset in terms of depth for 25% missing data*

3. Conclusion

The purpose of the correlation-based imputation procedure is to estimate the missing data as accurately as possible, thus that the reality of the refilled data matrix can be verified. The not-a-number values are used to fill in the missing data, then the calculation is performed on the dataset. In this paper, this type of procedure has been shown to be very effective in estimating missing data in the case where two columns are strongly correlated with each other regardless of the size of missing values. The completed dataset can be used for quantitative analysis of well logs to calculate the most important thermal reservoir parameters and lithological characteristics of the sequence of strata.

References

[1] Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* John Wiley & Sons, New York, 258 p. **https://doi.org/10.1002/bimj.4710310118**

[2] Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63 (3), pp. 581–592. **https://doi.org/10.1093/biomet/63.3.581**

[3] Máder, M. P. (2005). Módszertani Tanulmányok – Az imputálási eljárások hatékonysága. *Statisztikai Szemle*, 83 (7), pp. 628–643.

[4] Szabó, N. P., Nehéz, K., Hornyák, O., Piller, I., Deák, Cs., Hanzelik, P. P., Kutasi, Cs., Ott, K. (2019). Cluster analysis of core measurements using heterogeneous data sources: An application to complex Miocene reservoirs. *Journal of Petroleum Science and Engineering*, 178, pp. 575–585. **https://doi.org/10.1016/j.petrol.2019.03.067**

[5] Dong, Y., Peng, CY. J. (2013). Principled missing data methods for researchers. *Springerplus*, 2, p. 222. **https://doi.org/10.1186/2193-1801-2-222**

[6] Völgyi, L. (1984). A Nyírség potenciális szénhidrogénföldtana. *Földtani Közlöny*, 114 (2), pp. 161–169.

[7] Buró, B. (2015). *Recens és szubrecens felszínformáló folyamatok vizsgálata nyírségi mintaterületeken.* Egyetemi doktori (PhD-) értekezés, Debreceni Egyetem, 150 p.

[8] McDermit, M., Funk, R., Dennis, M. (1999). *Data cleaning and replacement of missing values.* (Kézirat.)

[9] Cool, A. L. (2000). *A review of methods for dealing with missing data*. Texas, A&M University, (Kézirat.)