Missing Data In The Oil Industry And Methods Of Imputations Using Spss: The Impact On Reserve Estimation

[1] Robert Kosova, [2]Evgjeni Xhafaj, [3]Altin Karriqi, [1]Blerina Boci, [1]Dorina Guxholli

Department of Mathematics. Faculty of Information Technology.
“A. Moisiu” University. Durres. Albania.
[2] Department of Mathematics. Faculty of Engineering Mathematics and Physics.
[3]Department of Earth Sciences. Faculty of Geology and Mining.
Polytechnic University of Tirana. Albania.
Email: romathsc@gmail.com

***Abstract*—Missing data in statistical studies can be a serious challenge for researchers who need the accuracy and reliability of their work. Standard statistical methods work on the premise that the information on which they analyze, and conclude is complete and sufficient. In any statistical study, missing data or observations, regardless of their number, will reduce the sample size and, as a result, the accuracy of applied statistical methods will be undermined, and the statistical power will be weakened. This article aims to present the essential definitions and concepts of missing data in a statistical study, then analyze the data provided by geological studies for reserve estimation of K/D oilfield in Albania, test if the data are MCAR and substitute them with the most appropriate values using the SPSS software. In addition, the impact of missing data on the oilfield reserves calculation with the volumetric formula of OOIP will be evaluated.**

Keywords: missing data, imputation methods, volumetric formula, SPSS, oil reserves

I. INTRODUCTION

Missing data in statistical research and academic studies should be treated seriously because almost all standard statistical methods have been applied assuming that the information obtained is complete for all variables included in the statistical analysis. Missing data can cause a significant reduction in sample size and, as a result, will undermine the accuracy of reliability intervals, weaken statistical power as well as possible bias in parameter estimates.

Proper analysis and treatment of data loss can be challenging as it requires a careful examination of the data to identify the cause and pattern of missing data, as well as to find the best imputation method to replace them.

Missing dates are usually attributed to human error in data processing, machine error due to equipment malfunction, respondents' refusal to answer certain questions, dropout, lost archives or data documents, and unrelated data aggregation. Missing data is created especially in family and social surveys and questionnaires where respondents may avoid giving some information that they consider personal, some questions may be incomprehensible, or the participants may simply have forgotten to answer.

In industrial studies, such as oil exploration or production, where data is taken from samples, laboratory analyses, daily production data, missing data can be caused by technical problems, performance errors, sampling problems, laboratory errors, loss of archives, datasheets, etc.

Missing values are endemic across the social sciences and family studies [1]. In political surveys about 50% of the participants' data have missing values, in social and family research the percentage of missing data often approximates this level of missing values; generally many of the major data sets that are utilized in articles appearing in family journals have serious problems with missing values [2]. It is estimated that a missing data rate of 15% to 20% is common in educational and psychological studies [3].

Data may be missing for various reasons, which may be objective or subjective. The reasons that data is missing can affect the appropriateness and value of the methods used to address the problem [4].

Some of the best analyses and treatments of missing data are by Little and Rubin (2002); Allison (2001); and Howell (2007), [5-7].

The problem of the missing values depends in part on the percentage of missing data, the missing data model, and the type of the missing data. The model, quantity, and mechanism of missing values have significant effects on the outcome of a study [8]. Before applying any imputation methods to replace missing data, the researcher must first diagnose and understand the missing data processes underlying the missing data [9].

Various estimates are made about the maximum percentage of missing data that can be neglected without causing seriously biased results. One estimate is that in statistical studies, if the amount of missing data is less than 5% of all the data, then the listwise method can be used.

Peng (2006) noted that in several quantitative studies based upon surveys or questionnaires, published in education and psychology journals during the period 1998 to 2004, 36% of the studies had no missing data, 48% had missing data, and about 16% could not be determined [10].
Furthermore, among the studies that presented and treated the missing data, 97% of them used the listwise deletion method or the pairwise deletion method to deal with missing data, which are well-known methods for biased and ineffective evaluations in most statistical studies [11].

The percentage of missing data is important because it is directly related to the results and the quality of statistical conclusions and it significantly affects the quality of the analysis and statistical interpretations.
Schafer (1997) estimates that a missing data rate of less than 5% of the data set is inconsequential; Bennet (2001) estimates that a missing data rate of more than 10% will produce biased statistical results [12-13].

However, the amount of missing data is not the only factor that affects the statistical results. Other factors such as the missing data mechanisms and the missing data patterns have a greater impact on statistical research results than the proportion of missing data [14].

Numerous articles and research studies have been conducted in various research fields, such as sociology (Rafteri, 2000), political science (King et al., 1998), psychology (Schlomer et al., 2010), education (Cheema, J. R., 2014), communications (Harel, Zimmerman, & Dekhtyar, 2008), oil industry (Albertoni, 2003), (Wang et al., 2019).
They have analyzed problems created by missing data and have implemented different methods of data imputation, including deterministic methods (simple calculation, mean, regression, etc.,) and probabilistic methods (value estimation) [15]-[21].

The problem of missing data has been mentioned and addressed in several research studies in different topics in Albania; in familiar, social, academic, Halidini et al., (2017), Xhafaj et al., (2021); in economic, tourism services, and industrial studies, Kosova et al., (2015; 2020), [22-25].

However, in academic research and studies, there is always the question of understanding the reason for missing data and what to do and how to deal with missing data how to assess its impact on the study results.

II. CATEGORIZATIONS OF MISSING DATA

Missing Completely at Random (MCAR)

Data is considered missing completely at random when the probability of whether an individual is missing a value on a given measurement is unpredictable. That is, there is no systematic underlying process as to why individuals are missing for a given measurement. It may be the case that some of the questionnaires were accidentally dropped for one or a few participants, or that another person was momentarily distracted by other things unrelated to the question or the answer or random technical errors. In this case, the missing data is unrelated to both the missing and observed values in the dataset.

$$p(R/D_{mis}, D_{obs}) = p(R)$$

Missing at Random (MAR)

Data is considered MAR if it is missing because of some potentially observable, non-random, systematic process. Missing data is MAR if the probability of missing data for some variable (Y) is predictable based on the value of the other variable (X) or set of variables. The missing data depends only on observed values in the dataset.

$$p(R/D_{mis}, D_{obs}) = p(R/D_{obs})$$

Missing Not at Random (MNAR)

The missing data is MNAR if data is missing due to the value of the variable being under consideration. If we were considering a variable Y, it would be MNAR if individuals chose not to respond because of the value of Y.

$$p(R/D_{mis}, D_{obs}) \neq p(R/D_{obs})$$

A classic example is the questionnaires about income. Income may often be MNAR because people who make a very high or very low income might choose not to report their income because they do not feel comfortable. In this case, the missing data of the income variable is dependent upon the value of the variable, figure 1, (red is missing data in the *y*-variable, and blue is observed data).

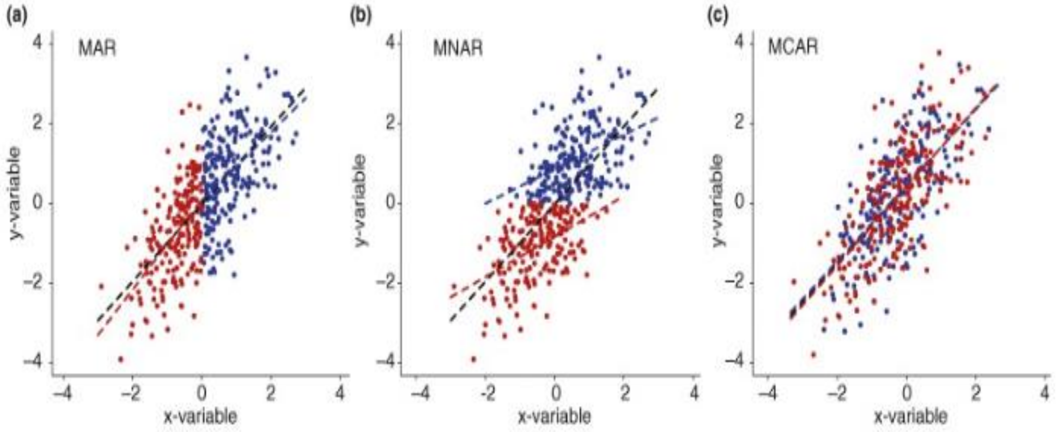

Fig. 1. *Illustration of the missing data classification.*
Source: Nakagawa & Freckleton (2008)

Methods of dealing with missing data:

Listwise deletion or case deleting

The most common method and the easiest way to deal with missing data that is used by most researchers is the list-wise method, which means deleting all the cases where there is at least one missing data point. Listwise deletion is an easy and simple method to implement because it is the default used in many statistical packages, including SPSS. The most obvious disadvantage of the listwise method is that it often deletes a large fraction of the sample, leading to a severe loss of statistical power, leading to larger standard errors, wider confidence intervals, and a loss of power in testing hypotheses. The listwise deletion method typically results in the loss of 20%–50% of the data [26].

If the data is MCAR then the listwise deletion will not introduce any bias into parameter estimates, because, under MCAR, the subsample of cases with complete data is equivalent to a simple random sample from the original target sample. It is also well known that simple random sampling does not cause bias.

Pairwise deletion

Pairwise deletion method deletes only the missing data in the present variables. In the end, some variables may have a different number of cases compared to other variables. Pairwise deletion is for linear models a very popular alternative replacement of the listwise deletion.

In such cases, the model is estimated using all available data for each variable or each pair of variables. Then, the sample statistics are substituted into the formulas for the estimation of the population parameters. In this way, all data are used and nothing is discarded. Like listwise deletion, however, if the data are MAR but not MCAR, the pairwise deletion may produce biased estimates. The pairwise deletion method may be more efficient than the listwise deletion method because more data are used in estimation formulas.

Imputation methods

Various methods fall under the general definition of imputation. The general definition is the method that produces some estimation for each missing value, by using the proper software.

Mean imputation is the simplest and the most popular imputation method for missing values, but it is well known to produce biased estimates. The mean is calculated for all the non-missing data variables and all the missing data will be substituted with the mean values. It has the advantage of keeping the same mean and the same sample size, and many disadvantages such as the false impression of sample size, and decreasing of the variance.

If the purpose of the research is mean estimations or if the data are missing completely at random, the series mean imputation will not bias the parameter estimate, it will still bias the standard error. On the other hand, mean imputation does not preserve the relationships among variables, the real relationship is quite underestimated [27]. Mean imputation usually leads to an underestimate of standard errors.

The mean of nearby points replaces missing values with the mean of surrounding present values (2 or more). Two or more valid values above or below the missing data are chosen to compute their mean to substitute the missing values.

Median of nearby points. Replaces missing values with the median of surrounding present values (2 or more). Two or more valid values above or below the missing data are chosen to compute the median to substitute the missing values.

Linear interpolation replaces missing values using linear regression produced from the present data using the method of least sum of squares. The last valid value before the missing value and the first valid value after the missing value are used for the interpolation.

The linear trend substitutes the missing values with the linear trend for that point. The method essentially performs a regression where the variable with missing values is the dependent variable and the case sequence number is the predictor. The existing series is regressed on an index variable scaled 1 to n and the missing values are replaced with their predicted values.

EM (Expectation-Maximization) method assumes a probability distribution for the missing data and estimates the likelihood under that distribution. Each iteration consists of an E and an M step. The E step finds the conditional expectation of the “missing” data, given the observed values and current estimates of the parameters, and then, these expectations are substituted for the “missing” data.

The Little’s chi-square statistic for testing whether values are missing completely at random (MCAR) or not is printed as a footnote to the EM matrix.

The null hypothesis is that the data are missing completely at random, and the p-value is significant at the 0.05 level.

If the p-value is less than 0.05, then the null hypothesis is refused, the data are not missing completely at random. In that case, the data may be missing at random (MAR) or not missing at random (NMAR). If the p-value is more than 0.05, then the null hypothesis is not refused, so the data are (MCAR).

III. MATERIALS AND METHODS

Missing data in the oil industry, especially in exploration and exploitation projects could be a common phenomenon because of the complex nature of the work. Even today, the majority of data is on paper, and there is a rich history in these documents that must be digitalized and used. Missing and corrupted data are among the main issues in the oil industry; they cost up to $60 billion annually [28].

The oil industry produces and uses a lot of data every day. Oil and gas exploitation performance contains data on daily production, such as oil and gas production rates (tons, barrels per day, m^3 gas per day, daily oilfield performance data such as pressure, temperature, flow, etc. Many other processes, such as oilfield exploration projects produce a lot of data, such as area, thickness, oil and water saturation, permeability, porosity, and other reservoir geological characteristics.

The problem of missing data could be a frequent occurrence in any oilfield exploration and exploitation project and dealing with them is a serious task. An incomplete dataset when is commonly simplified by ignoring all observations with missing values may lead to significant information loss. Traditional data imputation methods such as mean substitution, linear interpolation, and counting the most frequent values may produce bias in the data as the correlations between features are not considered. Thus, in such cases, probability methods and several multivariate imputation algorithms are considered better methods to estimate and replace the missing values [29].

The data collected from geological research are important for estimating the oilfield reserves. This means that the accuracy and completeness of the data may greatly affect the estimation of oil reserves, which is an important part of the projects. Different methods of missing data imputation may produce different results, which may affect the oil exploration projects, investments, net value expectations, and future work. A solution to such uncertainty may be better analyses of cause and the classification of the missing data, the revaluation of oilfield reserves, and finding the most appropriate methods of missing data imputation.

OOIP/IOIP and the volumetric formula

Oilfield Reservoir parameters that are used in the volumetric formula of OOIP/IOIP (Original/Initial Oil in Place) are:

Surface and thickness are the area and height of the oil-bearing layers included in the oil field. Each layer is considered as an independent oil field, and the total volume of the oilfield is the sum of the volumes of its layers.

Porosity is the percentage of pore volume or void space within reservoir rock that contains oil. Total porosity is the space in the rock whether or not it contributes to fluid flow. Effective porosity is the percentage of the rock interconnected pore volume that contributes to fluid flow in a reservoir. It excludes the rock isolated pores.

The ratio of average effective porosity to the total porosity is calculated into the volumetric formula of OOIP.

Pores with connection to other pores contribute to fluid movement in the reservoir; the higher the porosity of a formation, the more oil can be held in a given volume of rock. The porosity changes with burial

depth and usually declines with greater depths due to the compaction of the sediments.
A reservoir with very low porosity (less than 5 percent) has insignificant porosity, whilst excellent porosity is above 20 percent, table 1.

TABLE1. TYPICAL OIL RESERVOIR POROSITY VALUES

Porosity value, [%]	Classification
0-5	Insignificant
5-10	Poor
10-15	Fair
15-20	Good
>20	Excellent

Oil (gas, water) saturation is defined as a fraction, or percent, of the pore volume occupied by oil (gas, water). This property is expressed mathematically (for oil, gas, water), by the following relationships:

$$So = \{oil\ saturation\} = \frac{Oil\ in\ the\ pore\ volume}{pore\ volume}$$
$$Sg = \{gas\ saturation\} = \frac{Gas\ in\ the\ pore\ volume}{pore\ volume}$$
$$Sw = \{oil\ saturation\} = \frac{Water\ in\ the\ pore\ volume}{pore\, volume}$$

Reservoir rocks normally contain both hydrocarbon (oil and/or gas or only gas) and water. By definition, the sum of the saturations is 100%, therefore:

$$So + Sg + Sw = 1$$

Another important parameter connected to the porosity is the permeability which describes the ease with which a fluid can pass through the porous structure under a pressure drop. Porosity and permeability, which vary between reservoirs and even in the same reservoir, are the most important variables in characterizing and evaluating a reservoir.

Permeability is a measurement of a rock's ability, to transmit fluids. Formations that may transmit fluids such as sandstones are described as permeable.
Absolute permeability is the measurement of the permeability conducted when a single fluid, or phase, is present in the rock. Effective permeability is the ability to preferentially flow or transmit a particular fluid through a rock when other fluids are present in the reservoir (for example, effective permeability of gas in a gas-water reservoir).

Oil density is the ratio between the mass of oil produced to the volume. The density of crude oil can be determined from the specific gravity of the crude oil, solution gas gravity, solution gas-oil ratio, and oil formation volume factor (FVF).

The oil formation volume factor (FVF) is the volume of oil at natural conditions to the volume of oil at elevated pressure and temperature in the reservoir (bbl/STB- produced oil barrels to the stock oil volume in the reservoir). Values of (FVF) normally range from approximately 1.0 bbl/STB for crude oil systems containing little or no solution gas to nearly 3.0 bbl/STB for highly volatile oils.

Reserve estimation

Estimation of oilfield reserves is one of the most important tasks in oil and gas exploration projects.
Oil field reserves are quantities of oil and gas that are expected to be produced economically from discovered accumulations starting from a certain date and onwards. The discovered resources are the quantities of oil and gas that are estimated on a given date remaining in known accumulations plus the quantities already produced by these accumulations, (https://www.spe.org/en/industry/reserves/).

One of the main methods of estimating oil and gas reserves is the volumetric method. Volumetric methods involve the calculation of reservoir rock volume, the hydrocarbons in place in that rock volume, and the estimation of the portion of the hydrocarbons in place that ultimately will be recovered.

Parameters determining the volumetric reserves estimate are rock volume, which may simply be determined as the product of a drainage area and wellbore net pay or by more complex geological mapping, effective porosity, fluid saturation, and other reservoir parameters, and recovery factor (RF) which is the recoverable amount of hydrocarbon originally in place, normally expressed as a percentage.

For primary recovery, which is the first stage of hydrocarbon production, in which natural reservoir energy, such as gas drive, water drive, or gravity drainage, displaces hydrocarbons from the reservoir, into the wellbore and up to the surface, the RF is about 10% of oil originally in place (OOIP).

Reserves estimation is equal to the multiplication of oil is originally in place and recovery factor. R= OOIP* RF.
The volumetric formula to calculate the amount of Oil Originally in place (OOIP) is:

$$OOIP = A \ * h \ * \emptyset \ * So \ * Yn \ * \frac{1}{bn} \tag{1}$$

Where:
$OOIP$ – Oil Reserves (*tons, barrels, bbl*)
A – Oilfield area (m^2)
h –Average depth of reservoir (m),
$\emptyset$ – Porosity ratio (%).
So – Oil saturation (%)

γn –Oil density $\left(\frac{kg}{m^3}\right)$

bn –Formation Volume Factor

IV. RESULTS AND DISCUSSION

Data is provided by the K/D oil field, part of the Kuçova oil field, which was the first discovered oilfield in Albania, in 1928, [30].

The Kuçova oil field is a sandstone oilfield, and the second-largest oilfield in Albania, after the Patos-Marinza, which is the largest onshore oilfield in Europe. The Kuçova oil field is located near the city of Kuçova, in south-central Albania, 30 kilometers east of the city of Fier, figure 2.

The oil field consists of many layers of oil fields, which can be considered independent oil fields for the assessment of oil reserves, figure 3. The oil field is still active and produces about 1000 barrels per day, its proven reserves are about 490 million barrels, OOIP (IOIP) is estimated at 1 billion barrels.

The Kucova oilfield has produced more than 23 million barrels by the end of the year 2006. There are about 1000 active wells that are of relatively shallow depths (less than 1000 m) and with rather low productivity [31].

With a history of almost 100 years, there should have been a lot of data available for this oilfield and other oilfields of Albania. Missing data are related to daily and annual production as well as data related to the daily performance of production wells, as well as geological data of oil-bearing reservoirs.

This applies also to other Albanian oilfields. Needless to say, the oil industry in Albania which has a history of more than 100 years (since the first discovery in 1918) needs to have a complete, available database of all the oilfields, production data, and geologic data of the reservoirs.

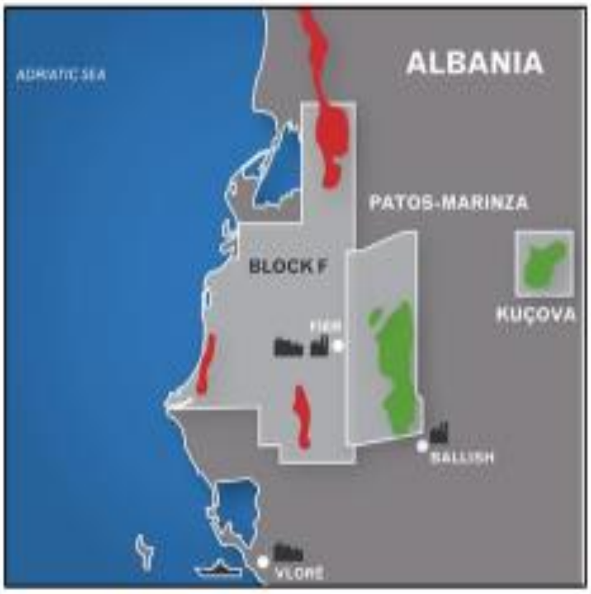

Fig. 2. *The Kucova and other oilfields of Albania*

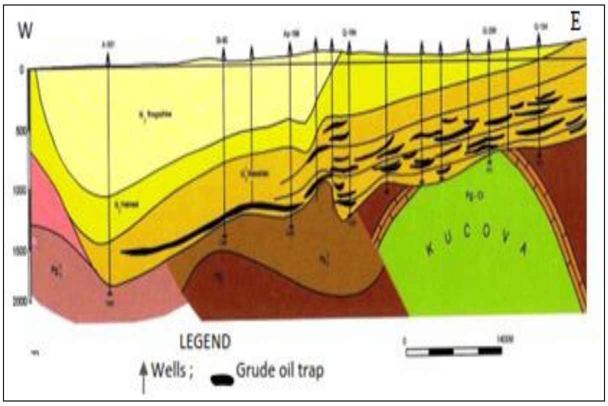

Fig. 3. *The Kucova oilfield with layers*

The full data matrix consists of 200 cases and 5 variables (parameters) for each case. The variables are layer area, depth, rock porosity, oil density, and oil saturation. Data processing is implemented with SPSS 24 program. A partial data is considered for this study, a 50 x 5 matrix. The parameters included in the volume formula are verified if their values are beyond the possible values (outliers). The first step in diagnosing data was to remove outliers because they may cause misinterpretation if not removed. The interval of values for the parameters of porosity, oil saturation, permeability, and oil density are well known for the present oilfield, as a result, the process of analyzing and cleaning the data showed no outliers. The same is for the area and thickness of layers, all the data are included in the known interval values.

Generally, the process of finding the outliers, which is included in the SPSS program is:

- Sort the data from low to high
- Identify the first quartile (Q1), the median, and the third quartile (Q3).
- Calculate the IQR = Q3 – Q1
- Calculate the upper fence = Q3 + (1.5 * IQR)

- Calculate the lower fence = Q1 – (1.5 * IQR)
- Use your fences to highlight any outliers, all values that fall outside your fences.

The SPSS produced only one extreme value for the Area parameter, which is not extreme for the real data, so it is considered in the formula.
The analyses of the data produced the percentage of missing values, by case, by variable, and overall and missing value patterns, figure 3-4.

The Little' test of MCAR (Missing Completely in Random) has produced the RESULTS (table 1).
The criteria of MCAR is satisfied Sig. = .632>.05, meaning that the missing data are Missing Completely at Random, (MCAR).
The univariate statistics, (mean, standard deviation, number and percentage of missing data and extreme values) descriptive statistics (number of available data) are provided by using the SPSS software. Also, the descriptive statistics, (minimum, maximum, mean, and standard deviation) are provided, table 2-3. The results of implementing the available SPPP imputation methods (listwise, available data, EM and Regression imputation method) the summary estimated means results are provided, table 4. For each parameter of the volumetric formula, all the available imputation methods are used (series mean, mean of nearby points (2), the median of nearby points (2), linear regression and linear trend, and listwise method). The only probability and most credible method is the EM method, and the data produced by that method are considered the most reliable method of imputation for the reserve estimation formula [32].
The OOIP formula considering all the data from included variables has produced the results, table 5. For the estimation of the OOIP, different values are calculated, minimum, maximum, Q25%, Q75%, and the average.
Three scenarios are estimated for the OOIP results, low, realistic, and high, which correspond to the Q25%, average, and Q75% value. For the results of OOIP, eight different values are provided, because of implementing all the imputation methods of SPSS and Excel 2016, too.

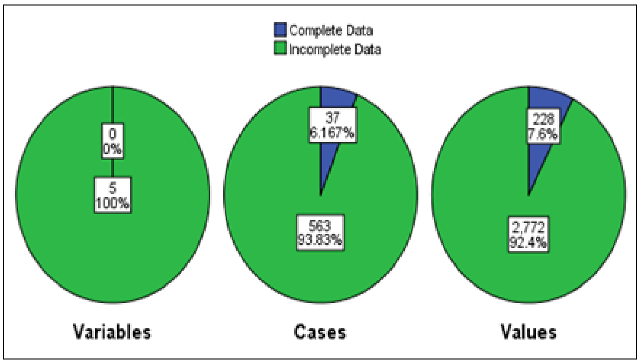

Fig. 4. *Overall summary of missing values*

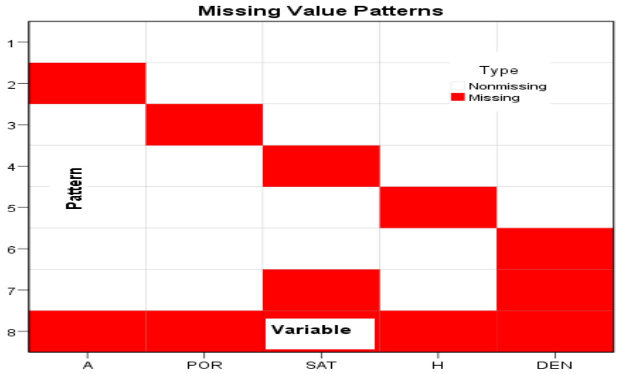

Fig 5. *Missing value patterns*

TABLE 1. LITTLE' TEST OF MCAR AND EM MEANS

EM Means ^{a,b}				
A	H	DEN	POR	SAT
2849617.820	390.1968	.8742	.1171	.3364
a. Little's MCAR test: Chi-Square = 20.161, DF = 23, Sig. = .632				
b. The EM algorithm failed to converge in 25 iterations.				

TABLE 2. UNIVARIATE STATISTICS OF AVAILABLE DATA

Univariate Statistics

	N	Mean	Std. Deviation	Missing		No. of Extremes [a]	
				Count	Percent	Low	High
A	47	2847184.7230	653650.88360	553	92.2	0	1
H	45	390.2667	120.25227	555	92.5	0	0
DEN	44	.8740	.03929	556	92.7	0	0
POR	46	.1178	.02732	554	92.3	0	0
SAT	46	.3380	.05784	554	92.3	0	0

a. Number of cases outside the range (Q1 - 1.5*IQR, Q3 + 1.5*IQR).

TABLE 3. DESCRIPTIVE STATISTICS

Descriptive Statistics

	N	Minimum	Maximum	Mean	Std. Deviation
A	47	1863332.00	4184568.0	2847184.7230	653650.9
H	45	245.00	625.00	390.2667	120.25227
DEN	44	.79	.93	.8740	.03929
POR	46	.06	.16	.1178	.02732
SAT	46	.25	.42	.3380	.05784
Valid N (listwise)	37				

TABLE 4. SUMMARY OF ESTIMATED MEANS

Summary of Estimated Means

	A	H	DEN	POR	SAT
Listwise	2773454.3780	396.0000	.8776	.1162	.3357
All Values	2847184.7230	390.2667	.8740	.1178	.3380
EM	2849617.8200	390.1968	.8742	.1171	.3364
Regression	2840236.3960	387.1085	.8741	.1187	.3353

TABLE 5. ESTIMATION OF OOIP OF THE PRESENT DATA AND IMPUTED

OOIP	Present data	Series mean	Nearby points	Median points	Linear regression	Linear trend	EM	Listwise
MIN	5382327.649	5409718.629	5409718.629	5409718.629	5409718.629	5409718.629	6799905.51	5382327.649
MAX	163976481.6	163449226.1	163449226.1	163449226.1	163449226.1	163449226.1	174524021.6	163976481.6
Q25	15259438.92	16618428.37	16295265.33	16240450.63	15048474.62	16618428.37	19414504.73	14104350.9
Q75	72804737.62	72645078.11	72645078.11	72645078.11	72645078.11	72645078.11	79272051	72844652.5
AV	38680426.99	38729890.84	38348942.41	38140558.79	39527383.05	38524862.66	38290761.71	37599613.46

V. CONCLUSION

The best solution for missing data is not to have one (Fisher, 1925).

Missing data are usually present in any statistical study, especially in the oil industry due to the complex nature of petroleum geology and other areas related to it.

The results of the study show differences in OOIP values calculated by different imputation methods. This means that the best thing for any study is working with complete and accurate data for reliable and credible results.

Considering the EM probability imputation method as the most reliable, the result shows that other methods produce different values, which may have a significant impact on the estimation of oil and gas reserves, for the three valuation scenarios.

The simplest way to fill in the missing data is to delete all incomplete cases. This method is included in most statistical programs such as SPSS, R, etc.

However, in cases where the missing data are MAR or NMAR or their percentage is more than 10%, this method will cause biased results and conclusions.

The oil industry in Albania has a history of more than 100 years and a complete database should have been created by now.

The collection of all the available oil industry data should be considered a valuable asset, which will help geologists to better analyze the situation of oil and gas fields in Albania, to make a reassessment of oil and gas reserves and resources, and better analyze their geological, petrographic characteristics, to discover mathematical models and patterns.

The purpose and realization of this study cannot exhaust the problem of treating missing data in the Albanian oil industry. It needs more complete studies and analysis, and much more data, in collaboration with geologists and other specialists in petroleum and mineral geology.

REFERENCES

[1] Juster, F., & Smith, J. (1998, August). Enhancing the quality of data on income and wealth: recent developments in survey methodology. In *25th General Conference of the International Association for Research in Income and Wealth. Cambridge, England*.

[2] Acock, A. C. (2005). Working with missing values. *Journal of Marriage and family*, *67*(4), 1012-1028.

[3] Enders, C. K. (2003). Using the expectation-maximization algorithm to estimate coefficient alpha for scales with item-level missing data. Psychological methods, 8(3), 322.

[4] Little, R. J., & Rubin, D. B. (1989). The analysis of social science data with missing values. Sociological Methods & Research, 18(2-3), 292-326.

[5] Little, R. J., & Rubin, D. B. (2002). Bayes and multiple imputations. *Statistical analysis with missing data*, 200-220.

[6] Allison, P. D. (2001). Missing data. Sage publications.

[7] Howell, D. C., (2007). The treatment of missing data. The Sage handbook of social science methodology, 208-224.

[8] Cook, R. M. (2021). Addressing missing data in quantitative counseling research. Counseling Outcome Research and Evaluation, 12(1), 43-53.

[9] Little, R. J., & Rubin, D. B. (2019). *Statistical analysis with missing data* (Vol. 793). John Wiley & Sons.

[10] Peng, C. Y. J., Harwell, M., Liou, S. M., & Ehman, L. H. (2006). Advances in missing data methods and implications for educational research. Real data analysis, 3178.

[11] Rubin, D.B. (1987). Multiple Imputation for Nonresponse in Surveys. J. Wiley & Sons, New York.

[12] Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. Chapman & Hall, London.

[13] Bennett, D. A. (2001). How can I deal with missing data in my study? Australian and New Zealand journal of public health, 25(5), 464-469.

[14] Lang, K. M., & Little, T. D. (2018). Principled missing data treatments. Prevention Science, 19(3), 284-294.

[15] Raftery, A. E. (2000). Statistics in sociology, 1950–2000. *Journal of the American Statistical Association*, *95*(450), 654-661.

[16] King, G., Honaker, J., Joseph, A., & Scheve, K. (1998, July). List-wise deletion is evil: what to do about missing data in political science. In Annual Meeting of the American Political Science Association, Boston (Vol. 52).

[17] Schlomer, G. L., Bauman, S., & Card, N. A. (2010). Best practices for missing data management in counseling psychology. Journal of Counseling Psychology, 57(1), 1.).

[18] Cheema, J. R. (2014). A review of missing data handling methods in education research. Review of Educational Research, 84(4), 487-508.

[19] Harel, O., Zimmerman, R., & Dekhtyar, O. (2008). Approaches to the handling of missing data in communication research. The Sage sourcebook of advanced data analysis methods for communication research, 349-371.

[20] Albertoni, A., & Lake, L. W. (2003). Inferring interwell connectivity only from well-rated fluctuations in waterfloods. SPE reservoir evaluation & engineering, 6(01), 6-16.

[21] Wang, M., Li, D., Qi, K., Xue, C., & Yang, E. (2019, October). SKNN Algorithm for Filling Missing Oil Data Based on KNN. In IOP Conference Series: Materials Science and Engineering (Vol. 612, No. 3, p. 032099). IOP Publishing.

[22] Halidini, D., Xhafaj, E., & Gjikaj, N. Treating the missing values for the total waste recycling in Albania. Interdisciplinary Journal of Research and

Development, Vol. 4, no. 1. 2017

[23] Xhafaj, E., Qendraj, D. H., Xhafaj, A., & Halidini, E. (2021). Analysis and Evaluation of Factors Affecting the Use of Google Classroom in Albania: A Partial Least Squares Structural Equation Modelling Approach. Math. & Stat., 9(2), 112-126.

[24] Kosova, R., Shehu, V., Naco, A., Xhafaj, E., Stana, A., & Ymeri, A. (2015). Monte Carlo simulation for estimating geologic oil reserves. A case study from Kucova Oilfield in Albania. Muzeul Olteniei Craiova. Oltenia. Studii şi comunicări. Ştiinţele Naturii, 31(2), 20-25.

[25] Kosova, R., & Sinaj, V. (2020). Service quality and hotel customer satisfaction: a case study from Durres, Albania. Annals of' Constantin Brancusi' University of Targu-Jiu. Economy Series, (6).

[26] King, G., Honaker, J., Joseph, A., & Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputations. American political science review, 95(1), 49-69.

[27] Patrician, P. A. (2002). Multiple imputations for missing data. Research in nursing & health, 25(1), 76-84.

[28] Nobakht, M., Mattar, L. et al., [2009] Diagnostics of Data Quality for Analysis of Production Data. Canadian International Petroleum Conference, Petroleum Society of Canada.

[29] Li, Y., Horne, R., Al Shmakhy, A., & Felix Menchaca, T. (2021, December). Reconstruction of Missing Segments in Well Data History Using Data Analytics. In Abu Dhabi International Petroleum Exhibition & Conference. OnePetro.

[30] URL: https://albpetrol.al/prodhimi-i-naftes-2/vendburimet/.

[31] Prifti, I., P., Nensi, M., & Suela, D. (2014). Petroleum system of the east part of the Ionian zone in Albania. *Online International Interdisciplinary Research Journal*, *4*(3), 49-65.

[32] Firat, M., Dikbas, F., Koç, A. C., & Gungor, M. (2010). Missing data analysis and homogeneity test for Turkish precipitation series. *Sadhana*, *35*(6), 707-720.

[33] Fisher, R.A. (1925). Statistical Methods for Research Workers. Oliver and Boyd (Edinburgh). ISBN 978-0-05-002170-5.

[34] Nakagawa, S., & Freckleton, R. P. (2008). Missing in action: the dangers of ignoring missing data. Trends in ecology & evolution, 23(11), 592-596.