Numerical Data Modelling and Classification in Marine Geology by the SPSS Statistics

Polina Lemenkova

Numerical Data Modelling and Classification in Marine Geology by the SPSS Statistics

Polina Lemenkova*‡

*College of Marine Geo-sciences, Ocean University of China, 238 Songling Rd., 266100, Qingdao, China

(pauline.lemenkova@gmail.com)

‡ Corresponding Author: Polina Lemenkova, 238 Songling Rd., 266100, Qingdao, China, Tel: +86-1768-554-1605, lemenkovapolina@stu.ouc.edu.cn or pauline.lemenkova@gmail.com

Received: 01.05.2019 Accepted:29.06.2019

Abstract- The paper focuses on the geostatistical analysis of the data set on the Philippine archipelago. The research question is understanding variability in several geospatial parameters (geology, geomorphology, tectonics and bathymetry) in different segments of the study area. The initial data set was generated in QGIS by digitizing 25 cross-sectioning profiles. The data set contained information on the geospatial parameters in the samples by profiles. Modelling and statistical analysis were performed in SPSS IBM Statistics software. The analysis of the topography shows strong variability of the elevations in the samples with the extreme depths in the central part of the study area (profile 13 with -9,400 m) and highest elevations in its south-western part (profile 17 with 1950 m). The analysis of the geological classes and lithology shows maximal samples of the basic volcanic rocks (40,40%) followed by mixed sedimentary consolidated rocks (31,90 %). Pairwise analysis of the sediment thickness and slope aspect demonstrates correlation between these two variables with the maximal sediment layer in the profiles 1-4 crossing the Philippines. The hierarchical dendrogram clustering of the bathymetry by three approaches shown maximal correlation of 5 clusters containing profile groups: 12-18 (centre), 22-25 (south-west), 1-2 (north), 7-8 (north-east), 19-21 (south-west). Other profiles show lesser similarities in the bathymetric patterns. The forecasting models were computed for the geospatial variables showing gradual increase in the gradient angles southwards and increased values for the sediment thickness in the north. Technically, the results proved effectiveness of the SPSS application of the geological data modelling.

Keywords Geological Modelling; SPSS Statistics; Data Analysis; Philippines; Marine Geology.

1. Introduction

Modelling geological data requires a multi-disciplinary approach to finding and analysing patterns in the categorical geospatial data. Understanding complex relationships between the variables constituting geospatial data set requires thorough analysis of the properties and features of the study area: geology, bathymetry, tectonics, geomorphology. A fusion of the different approaches and analytical methods and algorithms proposed by the existing statistical software and tools enables to solve such research problems and operate with the multi-source data.

Application of the data mining and machine learning algorithms for the geological domain is of particular interest nowadays, because the geospatial data are almost always include multidimensional data arrays collected through the geological expeditions as large data sets. Statistical analysis of the large data sets by various methods enables to perform precise and accurate data analysis, as well as perform predictive modelling, and visualize descriptive statistics on the raw data. Various approaches exist in data analysis and machine learning both covering general aspects of the data mining [1], [2], [3], [4], [5], as well as particular tasks of the geological modelling and data analysis in geosciences [6], [7], [8], [9], [10], [11].

The particular problem of the geoscience data modelling consists in traditional use of GIS and geoinformatics for the data analysis with significantly lesser usage of data analysis by the statistical algorithms and specific software (SPSS) or programming packages of R [12], [13], [14] or Python [15], [16], [17].

The data analysis in the geological marine mapping domains is traditionally based on the geoinformatics and embedded plugins of the GIS software [18], [19], [20]. Therefore, the statistical data analysis is often less

representative in the geosciences, comparing to the computer sciences, IT and finances. The statistical data analysis is widely used in other domain. These include, as mentioned before, such branches as economics, IT, finance and social studies where various algorithms and approaches are being developed, tested and improved. To fill in the gap between the geoscience and statistical data analysis, this research aims at the particular approach of the geodata modelling by means of the software SPSS IBM Statistics.

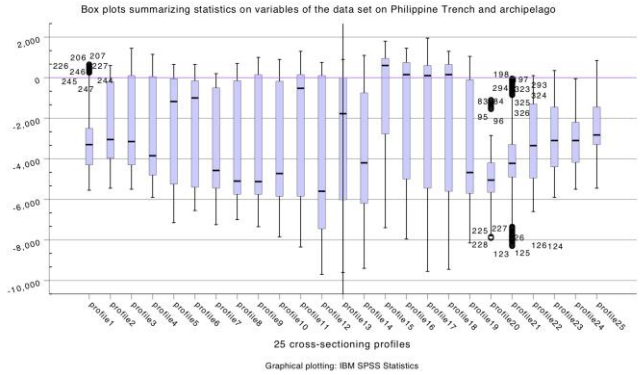


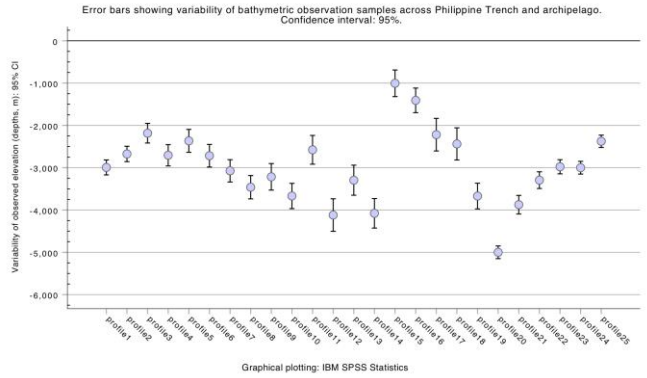
Fig. 1. Statistical box plots on the bathymetry: min-max values, 1st and 3rd quartiles, median.

The geological focus of the current research is analysis of the geological interrelationship between the variables (geomorphology, geology, tectonic plates, topography and bathymetry) of the study area: the Philippines islands and Philippine Trench located in the west Pacific Ocean, between the South China Sea and the Philippine Sea.

2. Methods

2.1. *Initializing GIS project and generating data set*

The first part of the research was performed in the Quantum GIS where the GIS project was initiated. The project contained vector thematic GIS layers with following



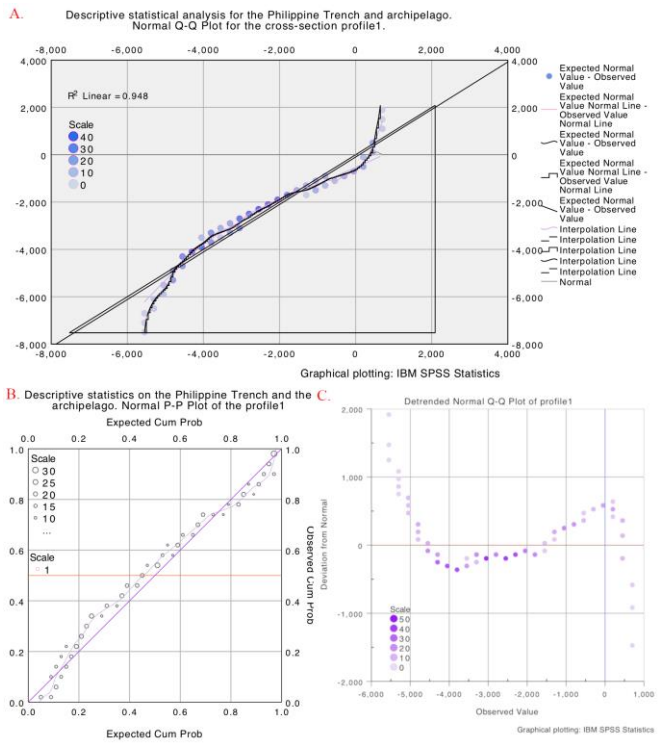
data: tectonics, geology, geomorphology, bathymetry.

Fig. 2. Error bar of variability (25 profiles) with a 5% confidence interval of the standard deviation.

The layers were processed, combined and re-projected in common coordinate system (UTM). Thereafter, the 25 cross-sectioning profiles were digitized across the Philippine archipelago and adjacent hadal trench crossing two tectonic plates: the Philippine Sea Plate and Sunda Plate. The information from all thematic layers where the digitized lines crossed the layers were recorded in the GIS attribute table which was then stored as a .csv table. At the next step, the table was imported int the IBM SPSS Statistics software as a working data set. All further research steps and graphical plotting were performed in the IBM SPSS Statistics.

2.2. *Analysis of the bathymetric ranges: variability of the samples by box plots and error bars*

Statistical analysis of the observations from the geospatial data set by visualizing box plots allows to distinguish concentration of the data showing the extreme values in the bathymetric and topographic elevations (Figure 1) across the Philippine Trench and archipelago. Plotted 25 box plots are constructed from five values each: minimum and maximum values, the first and third quartiles, and the median (Figure 1). The error bars for the bathymetric analysis were plotted using Graphs/Legacy Dialog menu from the SPSS. They show summaries for the separate



variables of 25 cross-sectioning profiles (Figure 2).

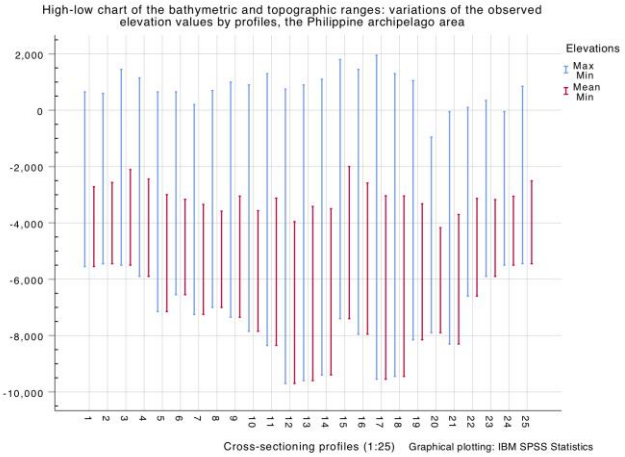
Fig. 3. Descriptive statistical analysis of the bathymetry by cross-sectioning profile 1. A) Normal QQ plot; B) Normal PP plot; C) Detrended normal QQ profile.

Each bar represents the confidence interval for the recorded means with excluded missing values showing the percentage value of 5% of the standard deviation. Descriptive statistical analysis on the bathymetry was performed by plotting the probability of the elevation data distribution by quantiles (QQ) and point-point (PP) plots. These methods show analysis of the normality of the

topographic data distribution. The PP plot shows a scatter diagram comparing two bathymetric samples of the same profile (Figure 3). The next step of the topographic analysis was performed by plotting statistical data description by the high-low charts for the analysis of the depths ranges showing summaries of the separate variables by the 25 profiles (Figure 4).

2.3. *Multi-variance analysis of the geospatial variables: geology, tectonics and geomorphology*

The analysis of the geological classes and lithology of the study area was performed by two approaches. First, qualitative analysis of the geological classes by the multi-plot visualization for comparative analysis (Figure 5). The chart shows mean values of the observation samples by separate variables (cross-sectioning profiles, 1:25 crossing the Philippine archipelago and trench. Second, the quantitative analysis of the geologic data (Figure 6) was performed using pie charts showing visual data distribution



across the Philippines.

Fig. 4. Statistical analysis by high-low charts for bathymetry and topography, the Philippines

Quantitative analysis of the tectonic values by the cross-sectioning profiles was performed by means of the drop line chart plotting aimed at the pairwise comparative analysis of the two tectonics plate: Sunda Plate and the Philippine Sea Plate (Figure 7).

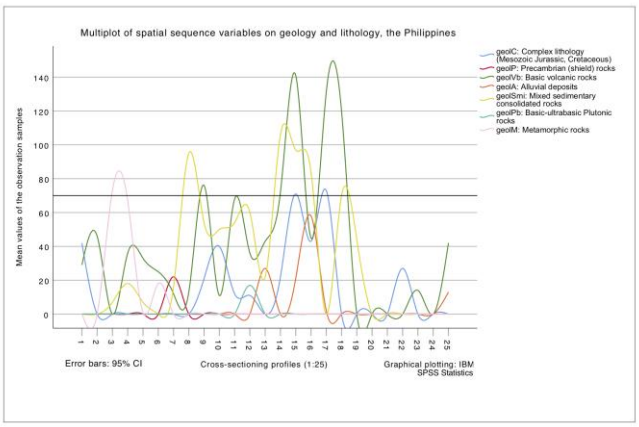
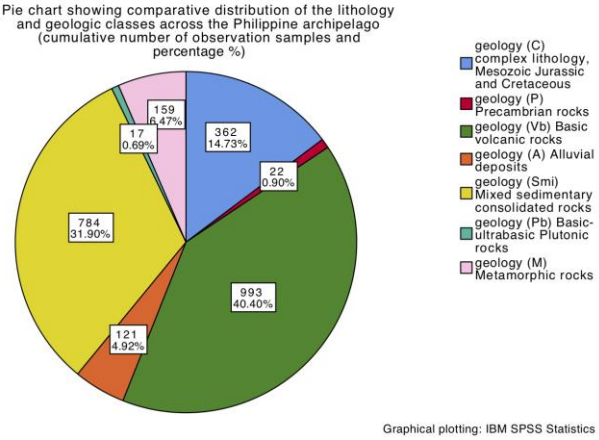


Fig. 5. Variation of the data distribution for the geologic classes across the 25 profiles

Pairwise correlation of the geology (sedimentation thickness layer) and geomorphology was done using method of the stacked area charts where two variables were visualized in a combined plot using relative y-scale factor (Figure 8).

2.4. *Hierarchical cluster analysis and modelling*

Visualizing data by the hierarchical cluster analysis and plotting dendrograms (tree diagrams) is an effective method reviewed and discussed in various existing research works [21], [22], [23]. Hierarchical clustering aims at grouping data by the attribute similarities (Figure 8 and Figure 9). The hierarchical cluster analysis was performed using 'Analyse / Classify / Hierarchical Cluster' module by the SPSS Statistics. The statistical methods included agglomeration schedule of the variables and computing proximity matrix



(Table 2).

Fig. 6. Quantitative data distribution on lithology and geology across the Philippine Islands

Table 1. Descriptive Statistics for the recorded geological classes across 25 profiles, Philippines

Class name	Maximum	Mean	Std. Deviation
geolC	73	14.48	22.602
geolP	22	.88	4.400
geolVb	141	39.72	41.390
geolA	58	4.84	13.104
geolSmi	106	31.36	36.555
geolPb	17	.68	3.400
geolM	73	6.36	19.506

Three methods of the data grouping were tested to achieve optimal results: Ward's method, Centroid clustering and Average linkage between groups. The best solution was proposed by the Ward's criterion method, a general agglomerative hierarchical clustering procedure, initially suggested by J. H. Ward [24] and developed in further studies [25], [26].

The particularity of this method consists in its approach towards data grouping: the criterion for choosing pair of clusters to merge at each step is based on the optimal value of an objective function. Modelling forecasts for the geological and geomorphic variables enables to asses the possible variations of the data using existing condition by given condition (Figure 11).

Table 2. Proximity Dissimilarity Matrix by Euclidean Distance. Geological variables: C, P, Vb, A, Smi, Pb, M

Geol.	geolC	geolP	geolVb	geolA	geolSmi	geolPb	geol M
C	.000	134.112	207.222	118.097	199.028	131.973	166.268
P	134.112	.000	283.748	72.049	237.563	27.803	103.087
Vb	207.222	283.748	.000	264.923	206.782	282.174	291.393
A	118.097	72.049	264.923	.000	215.307	70.682	121.861
Smi	199.028	237.563	206.782	215.307	.000	234.318	251.696
Pb	131.973	27.803	282.174	70.682	234.318	.000	102.137
M	166.268	103.087	291.393	121.861	251.696	102.137	.000

3. Results

The first part of the research (Figures 1-4) aims at the analysis of the bathymetry and topography. In this part the elevations across the study area were tested by various statistical approaches.

The results on the bathymetric analysis show following findings. As can be seen from the Figure 1, the greatest depths in bathymetry of the Philippine archipelago from the whole data set is noted by the profiles 12 and 13 where the values are reaching -9,150 and -9,400 meters, respectively. On the contrary, the highest elevation values are detected by the profiles Nr. 17 (1950 meters).

Error bars (Figure 2) demonstrate variability of the data (cross-sectioning 25 bathymetric profiles) with a 5% confidence interval of the standard deviation. As can be concluded by the Figure 2, the minimal error bars are recorded by the profile 20 with depth of -5,000m followed by the profile 24 and 25 with depths ranges of -3,000 and -2,400 meters, respectively. The PP plot (Figure 3) shows a scatter diagram comparing two bathymetric samples of the same profile. As can be seen from the Figure 3 B (lower left), the profiles Nr. 1 shows similar underlying distributions by the points with the line representing real data almost approaching modelled one, as shown by X axe (expected cumulative probability).

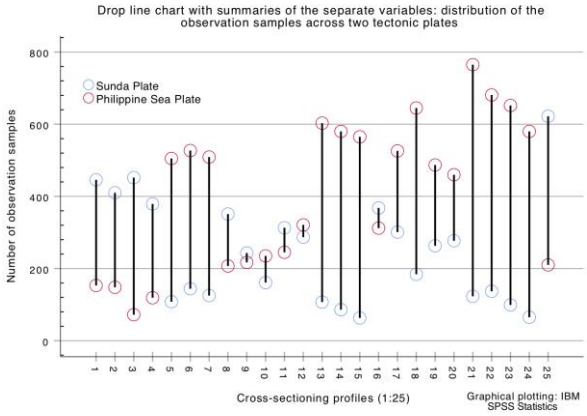


Fig. 7. Pairwise comparative analysis of the samples distribution across two tectonic plates

The probability of the sample points is visually expressed by the point size (Figure 3 B). Figure 3A shows example of the normal quantile-quantile (QQ) plot for the case of profile 1.

Table 3. ANOVA One-Way computing results for the geological variables

Geological variables		Sum of Squares	df	Mean Square
geolC	Between Groups, Total	12260.240	24	510.843
geolP	Between Groups, Tot.	464.640	24	19.360
geolVb	Between Groups, Tot.	41115.040	24	1713.127
geolA	Between Groups, Tot.	4121.360	24	171.723
geolSmi	Between Groups, Tot.	32069.760	24	1336.240
geolPb	Between Groups, Tot	277.440	24	11.560
geolM	Between Groups, Tot.	9131.760	24	380.490

The same data modelling was modelled for the 25 profiles of the data set. The data demonstrated normal

distribution. Figure 3C (lower right) shows a scatter diagram comparing a distribution of samples of the profile 1 by detrended normal QQ plot for this profile. Central part of the Philippine Trench (profiles Nr. 12-14) and south-eastern segment (profiles Nr. 17-19) are notable for the increased depth ranges, as can be seen from the Figure 4.

Table 4. Statistics on the frequencies in data records: sediment thickness and slope aspect (a). Calculated from grouped data. b). Multiple modes exist. The smallest value is shown. c). Percentiles are calculated from the grouped data.

Statistical values	Sediment thickness	Slope aspect
Mean	824.20	133.36
Std. Error of Mean	127.171	16.988
Median	629.00[a]	114.00[a]
Mode	402[b]	229
Std. Deviation	635.856	84.940
Variance	404312.250	7214.823
Skewness	2.946	.222
Std. Error Skewness	.464	.464
Kurtosis	8.088	-1.252
Std. Error of Kurtosis	.902	.902
Range	2504	277
Minimum	402	14
Maximum	2906	291
Sum	20605	3334

The depth variations in the bathymetric patterns are caused by the specific geomorphic conditions of the Philippine archipelago that crosses two tectonic plates and submarine volcanic areas.

Table 5. Model Statistics

Model	Model Fit statistics	Ljung-Box Q(18)		
	Stationary R-squared	Statistics	DF	Sig.
sedim_thickness-Model_1	.390	3.650	17	1.000
slope_aspect-Model_2	.251	12.269	17	.784
tan_angle-Model_3	.808	11.868	16	.753

The collision of the tectonic slabs directly affects the topography. Other factors include submarine erosions and geology (sedimentation and physical properties of rocks). On the contrary, the elevations are notable (Figure 4) for the profiles 15, 17 and 3 where the profiles cross the selected small islands from the Philippine archipelago.

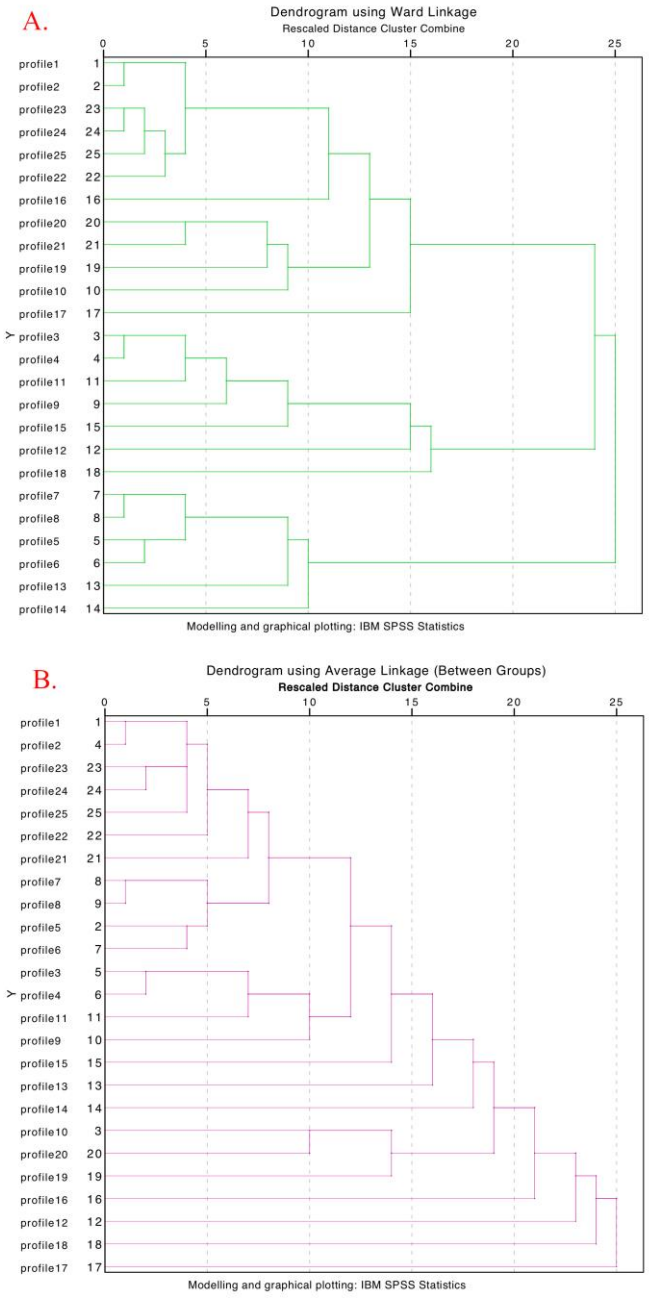


Fig. 8. Hierarchical cluster analysis with dendrogram visualization by two approaches: Ward Linkage and Average Linkage, for the set of cross-section profiles 1:25.

Analysis of the distribution of the lithologic classes by samples shows (Figure 5) the significant predominance of the 'Basic volcanic rocks' followed by the 'Mixed sedimentary consolidated rocks' formed by accumulating of the sediment particles, as the most representative of two classes recorded across the Philippine archipelago by cross-sectioning profiles.

The mixed sedimentary consolidated rocks were formed by the submarine erosion from the abyssal plain area located near the Philippine Trench and transported to the trench, as well as submarine canyons followed by the water currents. Metamorphic rocks formed as a result of the transformation of various types of the existing rocks created third large group of the lithological rock types on the Philippine archipelago.

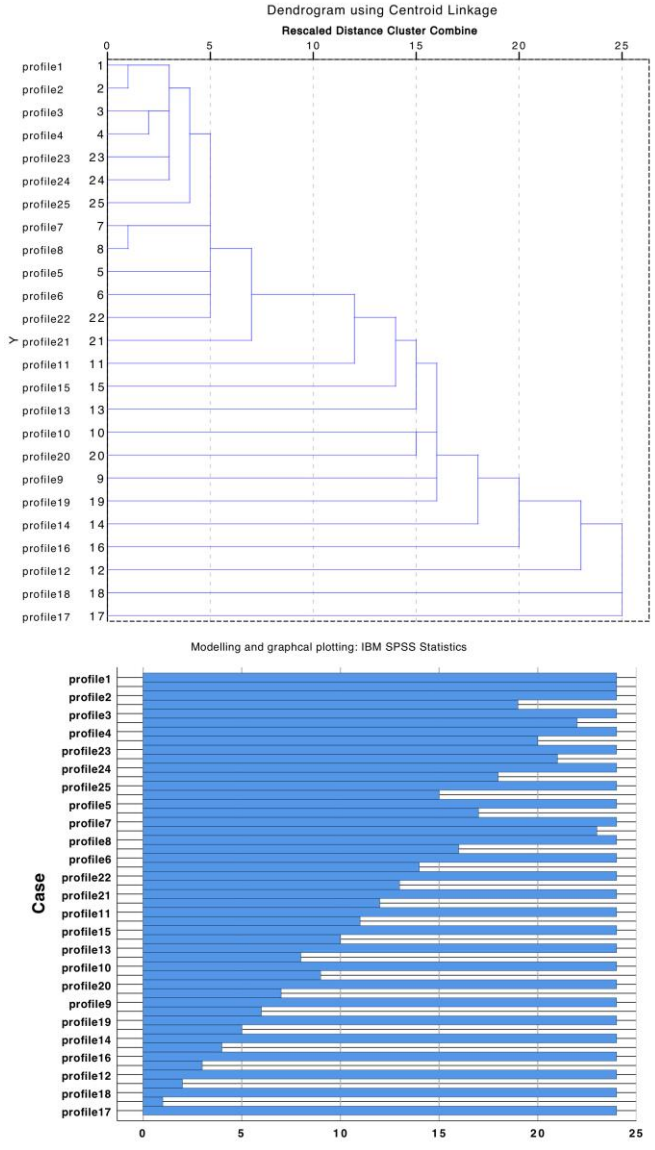


Fig. 9. Hierarchical cluster analysis with dendrogram visualization by Centroid Linkage approach and cases.

The analysis of the geological classes and lithology show maximal samples of the Basic volcanic rocks, Vb (40,40%) followed by Mixed sedimentary consolidated rocks, Smi (31,90 %) across the study area (Figure 6). On the contrary, Basic-ultrabasic plutonic rocks (Pb) are the least represented rock types among other lithological groups with the least samples recorded (0.69%).

This type of the rocks is presented by the igneous and meta-igneous rocks with a very low silica content, which constitute the lowerst part of oceanic crust of the Philippine Trench, generated at the mid-ocean or back-arc ridges in the area nearby. Precambrian rocks (P) are the second underrepresented rock type across the Philippines, as can be noticed on the Figure 6 (0.90%).

The Precambrian rocks are exceptionally rare since they are formed during the earliest geologic period of the Earth formation, and in most cases they are formed further to the metamorphic rocks. With 4,92 % and 6,47 % of the detected data samples, the Alluvial deposits (A) of the coastal area of the Philippines and the Metamorphic rocks (M) show the moderate part in the patters of the data distribution, respectively.

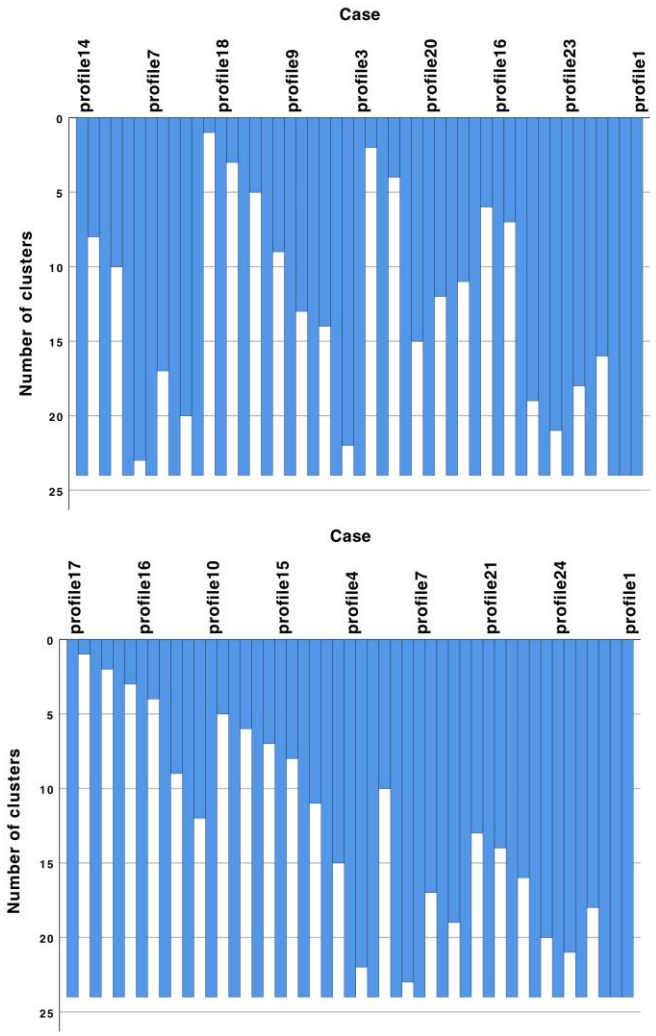


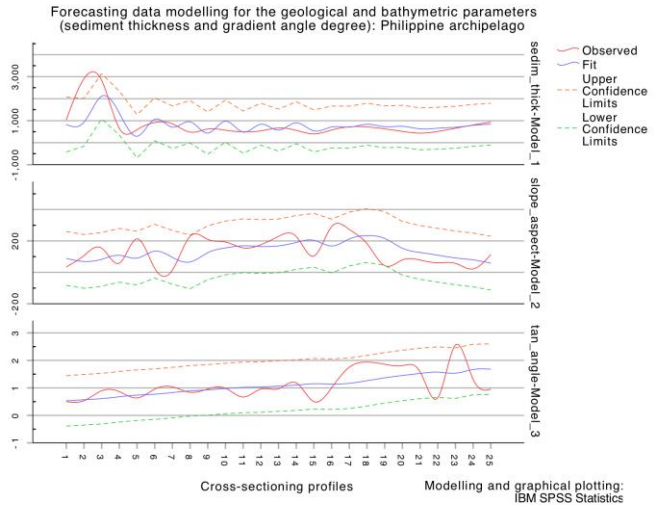
Fig. 10. Hierarchical cluster analysis: cluster cases.

Complex lithology of the Mesozoic, Jurassic and Cretaceous rocks is covered by the 14,73% of the total records in the observation samples.

The variations in the lithological patters of the Philippines may reflect the restricted geological and oceanographic conditions of the rocks formation, as well as the specific stratigraphy and sedimentary ocean environment of the Jurassic, Mesozoic and Cretaceous strata around the Philippines.

4. Results

The computation of the descriptive statistical parameters (standard deviation, maximal and mean values) is shown on the Table 1. The computation of the proximity dissimilarity matrix by Euclidean distance method is presented on the Table 2. It shows the two-dimensional square array containing the distances between the geological values across the seven recorded groups taken pairwise, between the



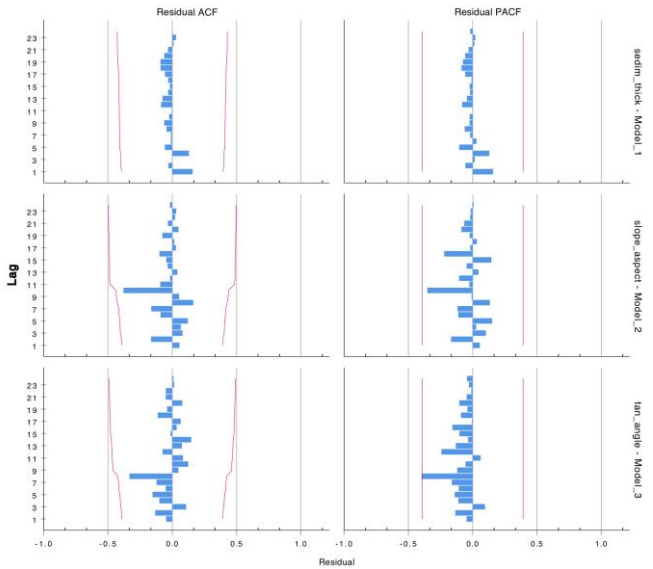
elements of the each set of the geological classes.

Fig. 11. Forecasting data modelling for the geospatial parameters: A) sediment thickness, slope aspect and gradient angles

The computations of the One-Way ANOVA for the geological modelling are shown on the Table 3 to performed to compare means of the seven geological samples. The graphical results of the drop line chart visualization, one of the experimental statistical methods, can be used to determine pairs of the factors that should be analysed comparatively. Two tectonic plates that cross the study area, Philippine Sea Plate and Sunda Plate, influence the geomorphology of the Philippines by the slab collision. The characteristics of the data distribution by two plates, respectively, can be used for the analysis of the spread of the tectonic plates in the study area. Thus, Figure 7 illustrates the maximal samples located on the Philippine Sea plate for the profiles 21, 22 and 23 (765, 681 and 652 samples, respectively), that is south-west location of the Philippines. Sunda Plate, a minor tectonic plate located mostly in the South-China Sea, has significantly lesser influence on the Philippine Trench bathymetry comparing to the Philippine Sea Plate. Thus, it shows much lesser observation samples with the maximal records for the profile 25 (622 samples).

The hierarchical dendrogram clustering of the bathymetry by three approaches shown maximal correlation of 5 clusters containing profile groups across the Philippines: 12-18 (centre), 22-25 (south-west), 1-2 (north), 7-8 (north-east), 19-21 (south-west). Other profiles show lesser similarities in the bathymetric patterns. Grouping bathymetric data by classes using hierarchical clustering (Figure 8) aimed at building agglomerative hierarchy of the bathymetric ranges divided by clusters according to the similarities in depths or topographic elevations for the segments located on the land areas (Philippine islands).

By comparing three types of the clustering, Ward, Average Linkage (Figure 8) and Centroid Linkage (Figure 9) for several cluster cases of the 25 profiles (Figure 10), the Ward approach demonstrates the most effective results of clustering approach dividing the data set into distinct three groups that are in turn subdivided into small groups of the profiles. The first class comprises profiles 1-2, 10, 16-17, 19-25 with similarities in the elevation ranges. The majority of the profiles of this class are located in the southern part of the tench (starting from 16th to 25th), and two profiles belongs to the northern part of the trench. The second class includes profiles 3-4, 9, 11-12, 15 and 18, mostly located in the central part of the crossing Philippine Trench. Finally, the third class includes the group of profiles 5-8 (north-eastern part of the area) and profiles 13-14 located in the central part of the trench. The division of the profiles is performed by the topographic similarity of the depth values, both in the absolute values and in their repeatability showing character



of the geomorphic patterns.

Fig. 12. Forecasting data modelling for the geospatial parameters: B) Residuals for the autocorrelation (ACF) and partial autocorrelation (PACF)

Final part of the research included forecasting models computed for the geospatial variables showing gradual increase in the gradient angles (in tangent) by 25 profiles in the southward direction and increased values for the sediment thickness for the profiles 1-4 (north). The computational part of this methodology is shown on the Table 5 (Statistical computations), Table 6 (Model fit) and Table 7 (Model fit: Percentile), respectively.

The forecasting data modelling shows following geospatial parameters: sediment thickness, slope aspect and gradient angles (Figure 11), residuals for the autocorrelation (ACF) and partial autocorrelation (PACF) (Figure 12). The autocorrelation of a geological parameters as a function of delay in the observation samples across the data samples (profiles 1:25). It shows similarity between the geologic observations as a function of the spatial lag between the samples in a range from 1 to 25 (Figure 12).

Table 6. Model fit

Fit Statistics	Mean	SE	Min	Max
Stationary R-squared	.483	.290	.251	.808
R-squared	.249	.179	.048	.390
RMSE	196.836	271.913	.444	507.186
MAPE	58.569	45.000	30.448	110.470
MaxAPE	278.281	264.341	106.147	582.646
MAE	120.275	151.231	.309	290.153
MaxAE	717.296	1095.299	1.021	1978.148
Normalized BIC	6.771	7.292	-1.365	12.715

In this way it is aimed to show repeating patterns and frequency in the geospatial observations of three types: sediment thickness layer, slope aspect, gradient angle degree.

Table 7. Model fit: Percentile

5	10	25	50	75	90	95
.251	.251	.251	.390	.808	.808	.808
.048	.048	.048	.307	.390	.390	.390
.444	.444	.444	82.878	507.186	507.186	507.186
30.448	30.448	30.448	34.790	110.470	110.470	110.470
106.147	106.147	106.147	146.050	582.646	582.646	582.646
.309	.309	.309	70.363	290.153	290.153	290.153
1.021	1.021	1.021	172.720	1978.148	1978.148	1978.148

General correlation between the sediment thickness and the slope gradient can be explained by the geomorphic properties of the hadal trench affecting the patterns of the sediment accumulation. Instability can be noticed by the profiles 1-4 where the correlation is not direct due to the geospatial location (crossing Philippine archipelago), followed by the stable pattern for the rest of the profiles. The increase in the sediment noticed for the profiles 1-4 shows the Philippine archipelago. The computations are presented on the Table 4.

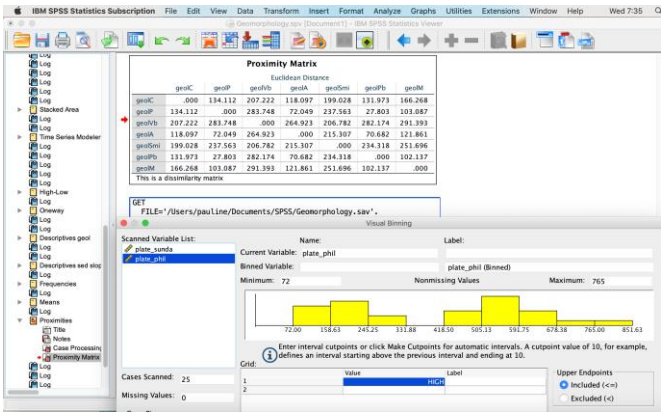


Fig. 13. Graphical User Interface of the project in IBM SPSS Statistics

5. Conclusion

The presented results show variations in the geomorphic, bathymetric and geological settings of the Philippine archipelago. The computations, modelling and visualization of the data was performed by the SPSS IBM Statistics software. Methodologically, the research demonstrated high effectiveness of the SPSS application towards the geological data modelling. Demonstrated multi-functional approach for the data processing is one of the defining tools in the geological data assessment. Compared to other tools, programs and methods, e.g. statistical libraries of R or Python programming languages, the SPSS IBM Statistics demonstrated more straightforwardness due to the GUI (graphical user interface), Figure 13. However, compared to the popular and widely used MS Excel, the SPSS IBM Statistics proposes much more functionalities for the advanced statistical modelling.

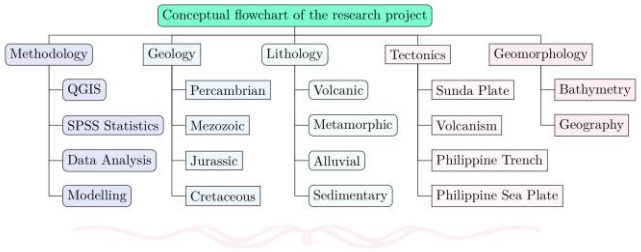


Fig. 14. Schematic methodological flowchart of the current research

Various statistical approaches have been tested to quantify the geological differences of the Philippine archipelago, as well as trends and variability in its bathymetric patterns and correlation between the geomorphic landforms and geological features (layers of slope aspect and sediment thickness). The use of different modelling algorithms, as well as the choice of the variables within the dataset to be modelled and used for numerical simulations can are important in the geological investigations [27], [28], [29]. Therefore, selecting statistical algorithms for the data processing is of high importance. The schematic research flowchart used in this research includes thematic clusters, such as geology, tectonics, geomorphology of the study area, as well as methodological approaches (GIS and statistical analysis), Figure 14.

This research is focused on the application of the existing tools, approaches and statistical modelling algorithms provided by IBM SPSS Statistics software towards geological modelling. Finding revealed how well tested variables (geomorphology, lithology, bathymetry) agree upon the geomorphic variations within the selected study area of the Philippines. The study furthermore demonstrated spatial variability of the geomorphology of the Philippine trench and adjusting archipelago. The consistent combination in various statistical metrics of the SPSS IBM Statistics shown a multi-factor analysis of the study area from different approaches (Figure 14): 1) bathymetry (assessment of variations in depth and elevations); 2) geology (assessment of various geologic classes of the Philippine islands); 3) geomorphology (assessment of the slope degree); 4) sedimentation (assessment of the variation in sediment thickness layer); 4) geometric variations of the hadal trench (assessment of the slope degree and aspect); 5) lithology (assessment of the properties and origin of the rocks); 6) tectonics (assessment of the distribution of the observation samples across two tectonic plates: Sunda Plate and the Philippine Sea Plate). All tested modelling and classification approaches performed by the SPSS on the data set and observation samples proved spatial variability in the geologic and geomorphic structure of the Philippine archipelago and hadal trench.

Acknowledgements

This research was funded by the China Scholarship Council (CSC), State Oceanic Administration (SOA), Marine Scholarship of China, Grant Nr. 2016SOA002, Beijing, People's Republic of China.

References

[1] A. S. Sidhu, C.Y. Cho, J.A. Leong, R.K.J. Tan, "Large Scale Data Analytics". Studies in Computational Intelligence Data, Semantics and Cloud Computing, vol. 806, pp. 89. Springer, Australia. doi: 10.1007/978-3-030-03892-2

[2] H. Cuesta, and S. Kumar. 2016. Practical Data Analysis, 2nd Edition. A practical guide to obtaining, transforming, exploring, and analyzing data using Python, MongoDB, and Apache Spark. pp. 360. ISBN-10: 1785289713. Packt Publishing Ltd. Livery Place, Birmingham, UK.

[3] P. Lemenkova. "R scripting libraries for comparative analysis of the correlation methods to identify factors affecting Mariana Trench formation". Journal of Marine Technolology and Environment, vol. 2, pp. 35-42, 2018. arXiv: 1812.01099, doi: 10.6084/m9.figshare.7434167

[4] C.D. Manning, P. Raghavan, and H. Schuetze, An introduction to information retrieval. Cambridge: Cambridge University Press, 2009.

[5] Y. Demchenko, P. Grosso, C. de Laat, P. Membrey, "Addressing big data issues in scientific data infrastructure," 2013 International Conference on Collaboration Technologies and Systems (CTS), San Diego, CA, 2013, pp. 48–55.

[6] J. Davis, Statistics and Data Analysis in Geology. Kansas Geological Survey John Wiley and Sons, 1990.

[7] F. Politz, B. Kazimi, and M. Sester, "Classification of Laser Scanning Data Using Deep Learning", vol. 38. Wissenschaftlich-Technische Jahrestagung der DGPF und PFGK18 Tagung in München – Publikationen der DGPF, Band 27, 2018.

[8] C. S. Campbell, P. W. Cleary, and M. Hopkins, "Large-scale landslide simulations: Global deformation, velocities and basal friction", Journal of Geophysical Research: Solid Earth, vol. 100(B5): pp. 8267–8283.

[9] P. Lemenkova, "Processing Oceanographic Data by Python Libraries Numpy, SciPy And Pandas", Aquatic Research, vol. 2(2), pp. 73-91, 2019, doi: 10.3153/AR19009

[10] S. H., Cannon, and W. Z. Savage, "A mass-change model for the estimation of debris-flow runout". The Journal of Geology, vol. 96(2), pp. 221–227, 1988.

[11] P. Lemenkova, 2018. "Factor Analysis by R Programming to Assess Variability Among Environmental Determinants of the Mariana Trench". Turkish Journal of Maritime and Marine Sciences, 4(2), pp. 146-155, doi: 10.6084/m9.figshare.7358207, 2018.

[12] R Development Core Team (2012). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, [Online] url: http://www.R-project.org/

[13] D. Sarkar, Lattice: Multivariate data visualization with R. pp.25, New York: Springer, 2008.

[14] P. Lemenkova, 2019. "An empirical study of R applications for data analysis in marine geology". Marine Science and Technology Bulletin, vol. 8(1): pp. 1–9, 2019. doi: 10.33714/masteb.486678

[15] G. van Rossum. Python Programming Language. 2011. [Online] url: https://www.python.org/

[16] I. Idris Python Data Analysis Learn how to apply powerful data analysis techniques with popular open source Python modules. 348 pp. Packt Publishing. Birmingham, UK, 2014. ISBN 978-1-78355-335-8.

[17] R. Johansson, Numerical Python. A Practical Techniques Approach for Industry. Urayasu, Chiba, Japan, 2015. doi: 10.1007/978-1-4842-0553-2

[18] L. Ferranti, S. Passaro, and G. de Alteriis. "Morphotectonics of the Gorringe Bank summit, eastern Atlantic Ocean, based on high-resolution multibeam bathymetry". Quaternary International, 332, 99-114, 2014. doi: 10.1016/j.quaint.2013.11.011

[19] J.T. Vázquez, B. Alonso, M.C. Fernández-Puga, M. Gómez-Ballesteros, J. Iglesias, D. Palomino, C. Roque, G. Ercilla, and V. Díaz-del-Río. "Seamounts along the Iberian Continental Margins". Boletín Geológico y Minero, vol. 126 (2-3), pp. 483-514, 2015.

[20] C. Yesson, R. C. Malcolm, M. L. Taylor, A. D. Rogers. 2011. "The global distribution of seamounts based on 30 arc seconds bathymetry data". Deep-Sea Research Part I: Oceanographic Research Papers, vol. 58, pp. 442-453. doi: 10.1016/j.dsr.2011.02.004

[21] Jain, A.K., and Dubes, R.C., Algorithms for Clustering Data, Englewood Cliffs NJ: Prentice-Hall, 1988.

[22] Meila, M., "Comparing clusterings – An information based distance". Journal of Multivariate Analysis, vol. 98(5), pp. 873–895, 2007.

[23] Kumaran, G., Allan, J., and McCallum, "A. Classification models for new event detection", International conference on information and knowledge management (CIKM2004). ACM, 2004.

[24] J.H. Ward, "Hierarchical Grouping to Optimize an Objective Function", Journal of the American Statistical Association, vol. 58, pp. 236–244, 1963.

[25] P. Lemenkova. "Hierarchical Cluster Analysis by R language for Pattern Recognition in the Bathymetric Data Frame: a Case Study of the Mariana Trench, Pacific Ocean", 5th International Conference Virtual Simulation, Prototyping and Industrial Design. Proceedings, Ed. M. N. Krasnyansky. Tambov, vol. 2 (5), pp. 147–152, Nov. 14–16, 2018., doi: 10.6084/m9.figshare.7531550

[26] Murtagh, F. "Ward's Hierarchical Agglomerative Clustering Method: Which Algorithms Implement Ward's Criterion?" Journal of Classification, vol. 31, pp. 274-295, 2014. doi: 10.1007/s00357-014-9161-z

[27] Gauer, P., A. Elverhoi, D. Issler, and F. V. De Blasio. 2006. On numerical simulations of subaqueous slides: back-calculations of laboratory experiments of clay-rich slides. Norsk Geologisk Tidsskrift, vol. 86(3), pp. 295.

[28] A. Cerioli, F. Torti, M. Riani. 2013. Algorithms from and for Nature and Life. Studies in Classification, Data Analysis, and Knowledge Organization. Eds: B. Lausen, D. V. d Poel, A. Ultsch. 547 pp. ISBN-10: 978-3-319-00034-3. Springer, doi: 10.1007/978-3-319-00035-0

[29] N. Boylan, C. Gaudin, D.J. White, and M.F. Randolph. "Modelling of submarine slides in the geotechnical centrifuge", 7th International Conference on Physical Modelling in Geotechnics (ICPMG), pp. 1095–1100. Zurich, Switzerland: ICPMG, 2010.

[30] J.M. Chambers. Software for Data Analysis Programming with R. Springer, pp. 237-288, 2008. doi: 10.1007/978-0-387-75936-4