Mar. Sci. Tech. Bull. (2019) 8(1): 1–9

e–ISSN: 2147–9666

info@masteb.com

http://dergipark.gov.tr/masteb

http://www.masteb.com/

DOI: 10.33714/masteb.486678

RESEARCH ARTICLE

An empirical study of R applications for data analysis in marine geology

Polina Lemenkova[1*]

[1] *Ocean University of China, College of Marine Geoscience, Qingdao, China*

ARTICLE INFO

Article History:

Received: 22.11.2018

Received in revised form: 27.02.2019

Accepted: 08.03.2019

Available online: 15.03.2019

Keywords:

R programming

Statistical analysis

Mariana Trench

Bathymetry

ABSTRACT

The study focuses on the application of R programming language towards marine geological research with a case study of Mariana Trench. Due to its logical and straightforward syntax, multi–functional standard libraries, R is especially attractive to the geologists for the scientific computing. Using R libraries, the unevenness of various factors affecting Mariana Trench geomorphic structure has been studied. These include sediment thickness, slope steepness, angle aspect, depth at the basement and magmatism of the nearby areas. Methods includes using following R libraries: {ggplot2} for regression analysis, Kernel density curves, compositional charts; {ggalt} for Dumbbell charts for data comparison by tectonic plates, ranking dot plots for correlation analysis; {vcd} for mosaic plots, silhouette plots for compositional similarities among the bathymetric profiles, association plots; {car} for ANOVA. Bathymetric GIS data processing was done in QGIS and LaTeX. The innovativeness of the work consists in the multi–disciplinary approach combining GIS analysis and statistical methods of R which contributes towards studies of ocean trenches, aimed at geospatial analysis of big data.

Please cite this paper as follows:

Lemenkova, P. (2019). An empirical study of R applications for data analysis in marine geology. *Marine Science and Technology Bulletin*, 8(1): 1–9.

Introduction

The Mariana Trench is the deepest point of the Earth located in the western part of the Pacific Ocean, eastwards to Philippine islands and China (Figure 1). It crosses four tectonic plates: Caroline, Pacific, Philippine Sea and Mariana. The Mariana Trench has unique features in its geomorphology, complex geological and lithological structures. The geologic features of the Mariana Trench are briefly discussed below.

Geographic Location

Mariana Trench belongs to the deepest trenches of the Earth, with maximal depths above 9–11 km, all of which are located in the western half of the Pacific Ocean.

* Corresponding author

E-mail address: lemenkovapolina@stu.ouc.edu.cn; pauline.lemenkova@gmail.com (P.Lemenkova)

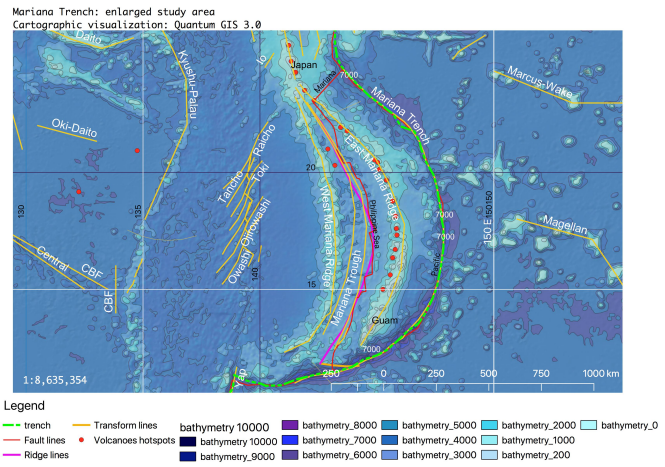

Figure 1. Study area: Mariana Trench (Cartography was created by Quantum GIS (QGIS) software)

The geographic location of the edge of the Mariana Trench along the coasts of the continents or island arcs is explained mainly by the subduction of oceanic tectonic plates at the boundaries of their collision with Mariana, Pacific, Philippine, Caroline tectonic plates (Ishibashi et al., 2015).

The seabed structure of the Mariana Trench is composed in the following way from top to bottom: a sedimentary cover, a basalt of the leitic composition, a complex of parallel dykes of diabases, an isotropic gabbro, a banded gabbro–ultrabasic complex lying on the mantle ultrabasites (Butuzova, 2003). There are differences in the velocity of the longitudinal seismic waves in the geological layers of the Mariana Trench structure: water and sedimentary layer have a speed of 3.5 – 6.2 km/s, basaltic layer has a speed of 6.5–7.0 km/s. The layer of gabbro and banded gabbro are located beneath. The boundary of the crust–mantle is made up by a sharp increase in the velocities from 7 to 8 km/s (the boundary of the Mohorovicic). The upper mantle is the region with the velocities of 8.0–8.2 km/s (Garfunkel et al., 1986). Analogues of the oceanic crust on the ground are ophiolites (Figure 2).

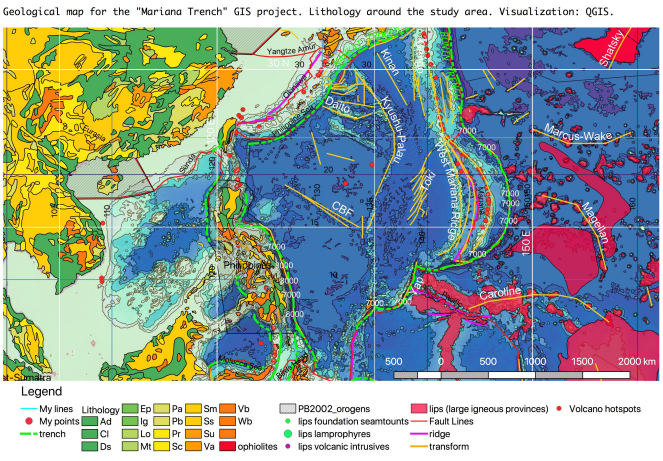

Figure 2. Lithology of the Mariana Trench area, Pacific Ocean (This map was created by QGIS software)

The average thickness of the crust of this type under the platforms is about 40 km (Gurevich, 1998). The regions with bark continental type in the ocean are fragments of continental mass formed as a result of the formation of the modern ocean floor causing migration and slow movement of the trench (Husson, 2012).

Sedimentation and Lithology

The lithological structure and sediment thickness are among the other factors affecting Mariana Trench formation. Relatively few other trenches, when compared to Mariana Trench, are also formed on the boundaries of the tectonic plates, but usually at a greater distance from the continents in the rift zones. Therefore, the sedimentation processes within Marina Trench are deeply impacted by the location near the rift zones that are mainly associated with the formation of the underwater mountain ranges and spreading. Their expansion in the sides of the neighbouring lithospheric plates is a result of the rise towards Mariana Trench.

Geometrically, the transversed faults are formed sub–orthogonal to major normal faults. They are created as local shortening structures with uneven depths above lateral and oblique extensional ramps. This creates excellent conditions for sedimentation accumulation of the outflow of the substance coming from the upper layer of the earth's mantle. According to available geophysical data (Dubinin and Ushakov, 2001), the oceanic crust of the tectonic plates surrounding Mariana Trench is composed by a number of layers. Therefore, the impact of the sediment subduction on the trench dynamics is relatively high, which is highlighted by Horleston and Helffrich (2012).

Geomorphology

The adjusting continental slopes are high, reaching up to several thousand meters and inclining to 3–6° (in the south–western part of the trench up to 30–40 °), the upper boundary of which coincides with the edges of the shelf (depths of 150–200 m). Slopes of the passive margins are strongly complicated by the terraces, ledges, marginal plateaus and canyons. Slopes of the margins are steeper, reaching can 5–7 km in height. The geomorphic structure of the Mariana Trench is complicated by the longitudinal ridges, steps, large landslide bodies and ledges. At the base of the continental slope of passive margins formed by the adjacent tectonic plates, Philippine Sea, Pacific and Caroline on the south–west, continental foot forms accumulative body. In turn, it is formed by the merged cones of the removal and plumes of the suspension flows and submarine landslides with abyssal sedimentation.

Submarine margins of continents or transition zones depend on the features of the relief and geology. According to them, they are divided into two types: the passive ones and the active ones. The first ones include the shelf that is mainland bank, the slope and the mainland foot. Having complicated relief structure with marginal seas, island arcs and deep–sea gutters, Mariana Trench belongs to the active type of trench. Due to the geodynamic reasons, Mariana Trench motion, migration and upper plate deformation can be described as the result of the response to the mantle flow and the imbalance between the forces exerted by the lower and upper continental plates at the plate interface.

Bathymetry

The deepest point of the Mariana Trench, the Challenger Deep, is located in the south–west of the trench with maximal depth of 10,984±25 m (95%) at 11.329903°N / 142.199305°E (Gardner et al., 2014). North–western Pacific Ocean is especially characteristic for the vast areas of the bottom of the basins occupied by depressions deeper 6000 m with a special case of Mariana Trench. Despite the extreme bathymetric values of the Mariana Trench at Challenger Deep, its structure has the following pattern: a major total area is being occupied by the abyssal depths (3–6 km), while the extreme depths exceeding 6 km, are smaller in comparison to the previous areas (Uyeda and Kanamori, 1979).

The abyssal make up relatively small part of the total area of the Mariana Trench, while the second ones cover about two thirds of the whole. Depths of more than 6000 m are confined mainly to the deepest part of the trench located in the south–western part of Mariana, although individual depressions to depths of 6–7 km, rarely up to 7.5 km, occur in the central part of the basins along the trench (Morgan, 1974). The typical depths of the ocean floor of the Mariana Trench are 4000–5000 m where mid–ocean ridges and deep–sea basins formed by numerous mountains and depressions are located. The general form of the bathymetric structure of the Mariana Trench is inclined toward the margin plate being connected with the ocean basic zone. It has high and steep island or continental slopes and more low and gentle slopes from the south–eastern side. The main area of the seafloor of the trench has prevailing depths of more than 3000 m, a seabed being the most ancient parts of the ocean floor formed in the late Jurassic.

Material and Methods

In this paper, a combination of various approaches has been used in the methodology workflow. Several R packages were applied for importing and manipulating geospatial data, combined with statistical machine learning. Various algorithms of data visualization were tested to facilitate spatial analysis using data from the GIS project: geology, tectonics, bathymetry, sedimentation, etc. The technical tools supporting this research include GIS, statistical methods and approaches made by R programming (e.g. Warner et al., 2008; Oliphant, 2007; Roberts et al., 2010). Recently developed powerful technologies provided by R, Matlab or Octava software, enable to perform precise computations and statistical analysis of big data as well as to create data frames in geosciences. Among others, a python language has become increasingly popular (Lin, 2008, Marta–Almeida et al., 2011). Nowadays, the machine learning and data mining for the oceanological research are among the most important technical goals. Using R programming (R Core Team, 2018) based machine learning algorithms ensure the preciseness and objectiveness of the big data set processing, which is always the case for oceanographic research.

However, a combination of the statistical methods with GIS dramatically increases the effectiveness of the research due to the embedded machine learning algorithms that enable to process big data frames dividing them into compatible data sets for statistical analysis. For instance, we can select several attribute tables specifically for magmatism of the nearby areas to analyse the possibilities of the earthquakes, or to focus on the bathymetric properties and distribution of various depth values across the selected profiles, etc. Special packages of the R programming include mathematical algorithms and a range of codes specially designed for statistical analysis, as demonstrated below. In the scope of current research, a functionality of R language has been tested. Indeed, it proved to be as an effective tool for studying distribution of the environmental factors affecting the structure of the trench, as well as bathymetric clusters grouped by the similarity of the geomorphic properties at the seafloor basement of the Mariana Trench.

The methodological workflow of the current work includes following steps:

i. Bathymetric GIS Data Processing in QGIS and LaTeX;
ii. Statistical analysis on data distribution: Kernel density and regression analysis with 3 approaches: linear, locally estimated scatterplot smoothing (loess) and general linear methods;
iii. Dumbbell charts for data pairwise comparison by tectonic plates;
iv. Ranking dot plots for correlation between the determinants (trench slope steepness in tg° angle by bathymetric profiles, igneous volcanic areas, depth);
v. Compositional "waffle" charts showing variation of the aspect and steepness classes of the trench basement angle, by 25 bathymetric profiles;
vi. Mosaic plots with standardised residuals to assess data distribution and silhouette plot for compositional similarities among the bathymetric profiles;
vii. Association plots with statistical Pearson residuals to analyse model fitness at each observation data set for generalized linear models;
viii. ANOVA (Analysis of Variance) to assess hypothesis testing.

Bathymetric GIS Data Processing

The digitizing of the 25 bathymetric profiles across the trench has been performed in QGIS.

The length of each profile was taken at 1000 km, and the distance between every pair was 100 km. The coordinates were saved in a table with three columns: elevations, latitude and longitude. After a square of the area was crossed by the profiles, the .csv table was imported for processing in LaTeX. The visualization of the 25 profiles grouped by three and four has been performed in LaTeX (Figure 3). Technically, the following script was used:

```
\begin{filecontents*}{MyTab18.csv}
ELEV,y2,x2
145.528246366,47.0433461696,–7800 # bathymetric data here in 3
    columns
\end{filecontents*}
\begin{tikzpicture}
```

\begin{axis}[grid=major,minor x tick num=10,minor y tick num=10,colorbar sampled line,colormap name=bluered,title={Mariana Trench. Bathymetric Profiles Nr.16,17,18},ylabel={Depth (m)},legend entries={Profile18,Profile17,Profile16,},scaled ticks=false,yticklabel style={

/pgf/number format/fixed,

/pgf/number format/fixed zerofill,}]

\addplot+ [scatter,only marks,mark=Mercedes star flipped,colormap name=bluered,] table [x=x, y=d, col sep=comma] {MyTab16.csv};

\addplot+ [scatter, colorbar sampled line,only marks,mark=asterisk,colormap name=bluered,] table [x=long, y=d, col sep=comma] {MyTab17.csv};

\addplot+ [scatter, colorbar sampled line,only marks,mark=10-pointed star,colormap name=bluered,] table [x=y2,y=ELEV, col sep=comma] {MyTab18.csv};

\end{axis} \end{tikzpicture}

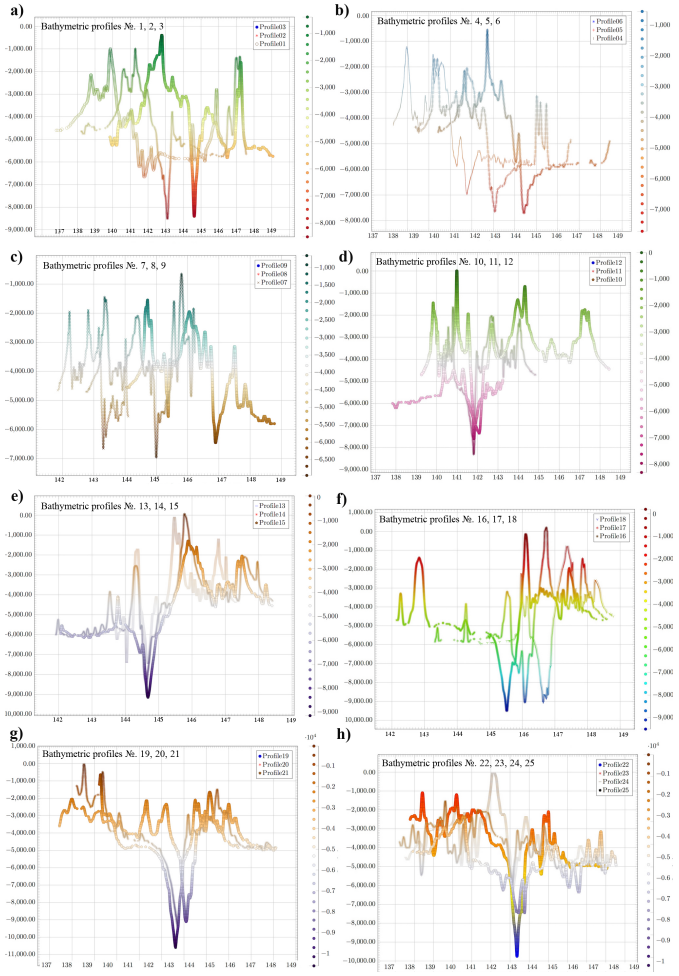

Figure 3. 25 cross-section bathymetric profiles of Mariana Trench. a) profiles 1, 2, 3; b) profiles 4, 5, 6; c) profiles 7, 8, 9; d) profiles 10, 11, 12; e) profiles 13, 14, 15; f) profiles 14, 15, 16; g) profiles 19, 20, 21; h) profiles 22, 23, 24, 25 (Plots were created in LaTeX)

Regression Analysis of Data Distribution: R library {ggplot2}

Hence, depending on the structure and size of the availability data frame, the methodological approaches of the assessment of data distribution may change. However, the most essential statistical analysis of the ocean research would start by the question of data distribution. Computing and plotting Kernel distribution curves, box plots, regression lines (probability of bathymetric data across profiles), quantile statistics, empirical distribution density function and other methods can be distinguished as the most useful for initial data processing.

The principle of the regression analysis is based on the analysis of the probability of the data values according to their actual distribution. A regression analysis represents outlying depths observations by each bathymetric profile, as shown on the example of three profiles in Figure 4. This methodology utilizes the relation between two or more quantitative depth variables so that one variable can be predicted from another. Thus, one can estimate the probability that the depth values will be located in this or that interval of values profile by profile according to the machine-based algorithm of the probability of their distribution. Kernel density distribution shows the majority of the depths of the trench between 3000 and 6000 m. The graph has been made as a combined plot (Figure 5) enabling to compare the overlapping and maximal aptitude of the density curves for both the profiles and depth data distribution by four tectonic plates.

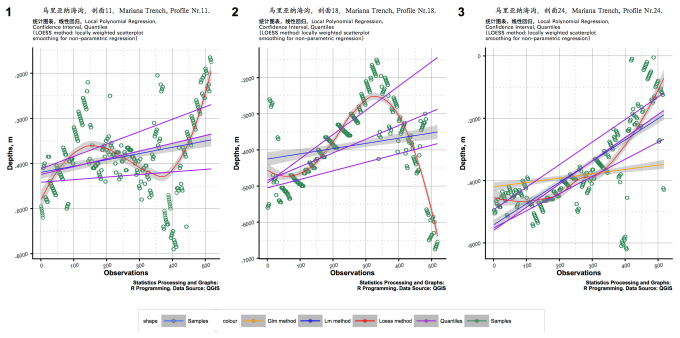

Figure 4. Regression analysis for selected profiles, R programming

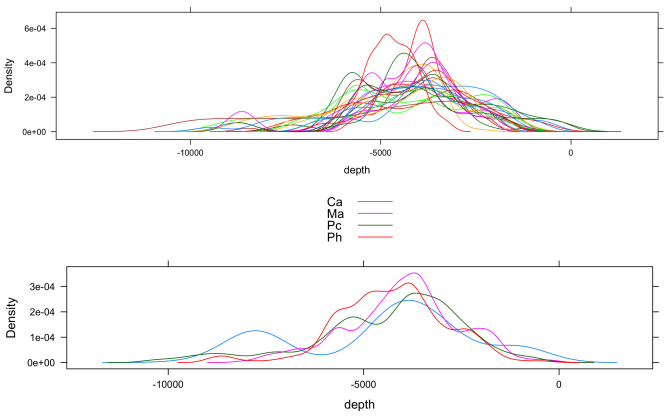

Figure 5. Kernel density estimation for 25 bathymetric profiles (Graph was created by using R)

Dumbbell Charts for Data Pairwise Comparison by Tectonic Plates

Because Mariana Trench crosses four tectonic plates (Mariana, Caroline, Philippine and Pacific) the comparative testing of the values distribution by adjacent tectonic plates has been completed through the Dumbbell chart plotting. Dumbbell chart is a visualization aimed to give an insight of how the margin tectonic plates constitute to the Mariana Trench pairwise. It is one of the new statistical methods which was initially widely used in biosciences, yet it can be applied to spatial analyses when it is necessary to take pairwise comparison of the data distribution from any thematic layer. Plotting has been carried out by calling R libraries {ggplot2} and {ggalt}. Dumbbell chart demonstrates (Figure 6) pairwise distribution of the bathymetric points constituting the continental plates, to show the composition of the sedimentary coverage by four tectonic plates.

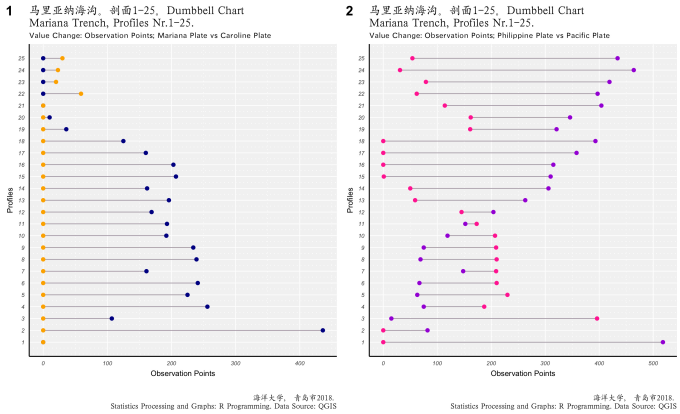

Figure 6. Dumbbell diagrams for pairwise comparison of the tectonic plates

The Y axis shows the profiles and the X axis – the number of observation points crossing to the given tectonic plate. In this case the comparison was given by pairs Philippine sea and Pacific plate, and Mariana and Caroline Plates, respectively. Although a complete comparison of all the environmental determinants has not yet been given (in the scope of this research only bathymetric points were plotted for the Dumbbell chart), a certain insight of the tectonic structure of the Mariana Trench contributes to understanding of the congruence of the ocean floor by four plates.

Ranking Dot Plots by Data Grouping

Pairwise correlation was visualized by the ranking dot plots, which is an effective tool to perform data grouping by variables: tectonic plates. Plotting distribution of the observation points by igneous volcanic areas aimed at visualizing areas affected by the magmatism. Large igneous volcanic areas contribute towards localization of possible earthquake zones. Thus, as can be drawn from the plot (Figure 7), the volcanic areas increase along Mariana Trench south–westwards.

Alike to the volcanic areas plotting, the variation of the steepness slope was performed using ranking dot plot: the bathymetric profiles with the steepest angles across the trench are located in the north–eastern part with a slight decrease towards the south–west. To calculate tgº angle of the profiles, a standardized formula was used, that is a relation of the maximal depth by profiles divided by the width of the corresponding bathymetric profile. The crucial points (Figure 7) were selected using library{ggalt} by calling following R code:

```
crucial_igneous<MDF[MDF$igneous_volc > 50 &
    MDF$igneous_volc <= 300 & MDF$profile > 5 &
    MDF$profile <= 25, ]

crucial_angles <- MDF[MDF$tg_angle > 0.00 & MDF$tg_angle
    <= 0.075 & MDF$profile > 5 & MDF$profile <= 22, ]
```

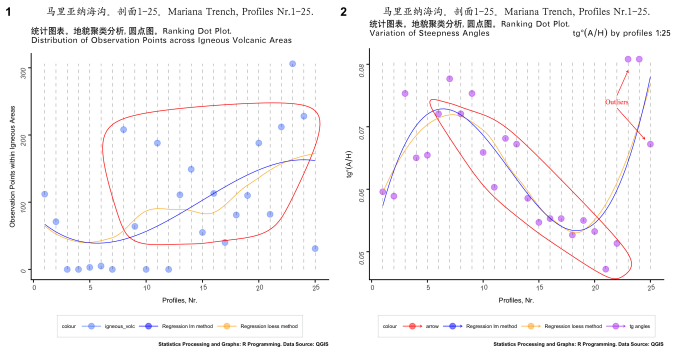

Figure 7. Ranking dot plots by data grouping

The distribution of angle steepness values across the trench have been shaped by the components of the igneous volcanic areas near the profiles, as understood from the Figure 7: profiles # 20, 22, 24, 8 and 11 with notable amount of igneous volcanic observation points (over 180) correlate with steepness angle.

Compositional Charts of the Determinants Variations

The constitution of the system composition has been visualized by the categorical plotting. One of the methods enabling categorical plotting is presented by compositional charts, sometimes referred to as "waffle charts". The core idea of this method is to demonstrate the division of the whole system by parts in percentage. Several R libraries can be used to perform technically compositional plotting, of which the {ggplot2} has been selected as the most effective: it enables to control the appearance of the plot by adjusting details (colors, font size and types, plotting two graphs together, etc).

The compositional chart is aimed to compare the distribution of the data across the study area, as well as the composition of the aspect class and slope angle for the Mariana Trench. As can be drawn from the chart (Figure 8), the category "very steep slope" is the dominant among all other geomorphological types of the slope degree of the Mariana Trench. Likewise, north and south–west aspect of the slope direction perfectly describes the geometry of the northern and southern part of the trench, respectively.

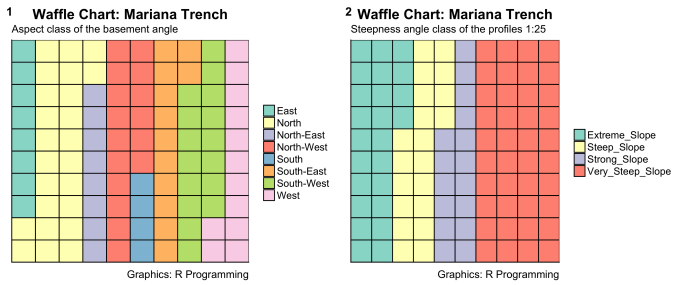

Figure 8. Compositional chart of geometric properties of the Mariana Trench

Mosaic and Silhouette Plots of the Observation Data by Tectonic Plates

Mosaic chart aims at categorical comparison of the geomorphological features of the trench by four tectonic plates. Visualizing mosaic plot (Figure 9) is a statistical method that involves a subdivision of a rectangular tile into the areas that represent the conditional relative frequency for a cell in the contingency table.

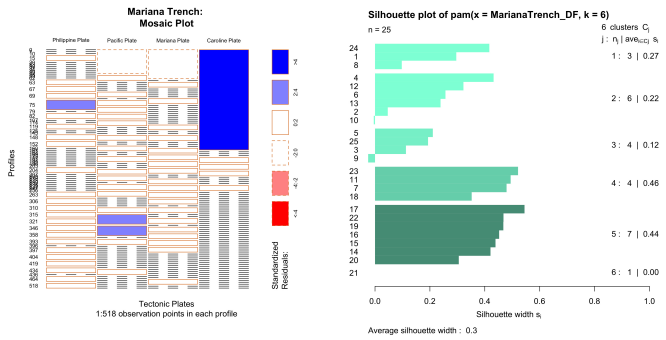

Figure 9. Mosaic plot and silhouette plots for categorical values of the environmental determinants

The algorithms and approaches used in mosaic plotting vary slightly. Upon examination of possible packages, the R library{vcd} was selected for this research. Using {vcd} library, each tile is colored to show the deviation from the expected frequency (residuals) from a Pearson X^2 or likelihood ratio G^2 test. The algorithm set includes execution of the following code:

MosaicTectRes<– mosaicplot(count, main = "Mariana Trench: \nMosaic Plot", sub = "1:518 observation points in each profile", xlab = "Tectonic Plates", ylab = "Profiles", las = 1, border = "chocolate", shade = TRUE)

The convenience of the mosaic plot for the geomorphological analysis of the ocean trenches consists in its visual representation of the association between the environmental variables. Thus, it gives an overview of the data structure and enables to recognize relationships between the different environmental variables. As can be seen from the Figure 9, the independence of the Caroline tectonic plate is shown clearly. Conversely, the boxes across categories constituting Pacific and Mariana plates have similar areas. The area of the tiles, the bin size, includes the identification of the sampling data (518 observation points across 25 bathymetric profiles), giving the proportional value to the number of observations within that category.

The interpretation and validation of consistency within clusters of geomorphologic and geological variables have been tested using silhouette method (Figure 9). This technique provides a succinct graphical representation of the suitability and fitness of each from 518 observation points across 25 bathymetric profiles within the clusters. The silhouette value measured the similarity of the points derived from the thematic layers (sediment thickness, slope steepness of the trench) to the cluster by cohesion and comparing them with other clusters, and separating them from the distinct points.

Thus, lower sediment thickness area would lie within one class while a class of areas with high level of sediment level fits to another. The silhouette ranges from −1 to +1 (Figure 9), where a high value indicates that the areas of trench are well matched to their clusters and poorly matched to the neighboring ones. To achieve the sufficient and appropriate configuration of the chosen cluster plotting, the geomorphic groups were grouped by values. The silhouette was calculated with Euclidean distance metric by calling library{cluster} using following R code:

pr5 <– pam(MarianaTrench_DF, 6)

str(si <– silhouette(pr5))

plot(si, col = c("aquamarine", "aquamarine1", "aquamarine2", "aquamarine3", "aquamarine4", "aquamarine4"))

The given script plots the silhouettes, returns the values in the n–by–1 vectors.

Association Plots with Statistical Pearson Residuals

A deviation from the independence of rows and columns containing bathymetric data and environmental variables is presented in a two–dimensional contingency table of the Cohen–Friendly association plots (Figure 10).

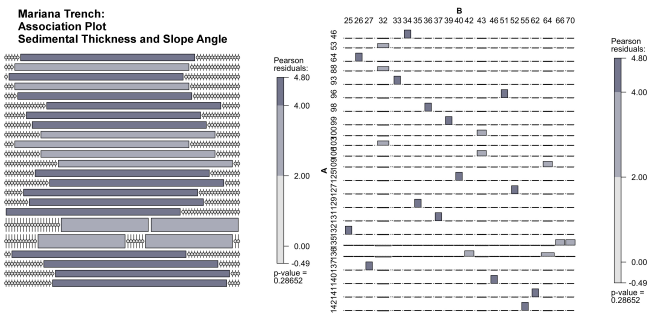

Figure 10. Association plot with statistical Pearson residuals, Mariana Trench

Extended association plot in R reveals relationships between the variables by the assoc() function in the {vcd} package. In this case, it is tectonic plates and geomorphological values (bathymetric depths, slope steepness). The Pearson's residuals on the association plot report test for the normality of residuals of the bathymetric observations. Computing Pearson's residuals enables to test whether the bathymetric samples are drawn from the identical distributions or a specified distribution, following the rules of the bathymetric structure. The

algorithm of the Pearson's computing represents the sum of the differences between the observed and expected outcome frequencies: the counts of the bathymetric observations along each of the 25 profiles. Each count in turn is squared and divided by the expectation. In this case, it enables to assess the model fit at each of 518 observations for every bathymetric profile for the generalized linear models within the context of the chi–square (Figure 10). It has been computed as a raw residual divided by the square root of the variance function. Technically, association plot method includes execution of the following R code:

library(vcd) # calling necessary package

MDF <– read.csv("Morphology.csv", header=TRUE, sep = ",") # uploading table

MDF <– na.omit(MDF) # deleting non available values in the data frame

row.has.na <– apply(MDF, 1, function(x){any(is.na(x))})

sum(row.has.na) # checking up non available values in the data frame

head(MDF) # visualizing data frame as a table

 # step–2. Merging columns in a table according to their category values (tectonic plates)

MDTt = melt(setDT(MDF), measure = patterns("^plate"), value.name = c("tectonics"))

levels(MDTt$variable) = c("Philippine Plate", "Pacific Plate", "Mariana Plate", "Caroline Plate")

assoc(count, shade=TRUE) # plotting association plot according to the tectonic plates crossing the Mariana Trench.

Analysis of Variance (ANOVA) to Assess Hypothesis Testing

The results have been interpreted on the homogeneity of variance by means of the one–way ANOVA series of tests (Figure 11). Several test methods were executed in this research. The brief description of their advantages and particularities is given below. The Tukey HSD (Tukey Honest Significant Differences) multiple pairwise–comparisons were applied to the ANOVA results. Since the ANOVA test is significant, the Tukey HSD test was computed by R function. TukeyHSD() takes the fitted ANOVA as an argument. This test enabled to perform multiple pairwise–comparison between the means of groups. The library{multcomp} was used to perform a pairwise t–test and to calculate pairwise comparisons between the group levels with the corrections for the multiple testing using R function pairwise.t.test(). The Levene's test was used to analyse the homogeneity of variances at the next step of ANOVA testing by calling function leveneTest() in the {car} package. Comparing to the pairwise t–test, the advantages of the Levene's test is its lesser sensitiveness towards the departures from the normal distribution.

Afterwards, the Welch test was executed in parallel to the oneway.test() in experimental mode. The Welch–test does not require that assumption have been implemented in the function oneway.test() which makes is more suitable comparing to the last one. The Shapiro test was alternatively used to analyse variance of the residuals (Figure 11, upper left). Finally, a non–parametric Kruskal–Wallis rank sum test was applied to finish with a series of ANOVA testing. The results have been interpreted on homogeneity of variance by means of one–way ANOVA tests. Since the p–value (Figure 11) is less than the significance level 0.05, one can conclude that there are significant differences between the groups highlighted with "*" in the model summary, that is four tectonic plates: Philippine, Pacific, Mariana and Caroline. The Tukey multiple pairwise–comparisons test enabled to perform multiple pairwise–comparison between the means of groups and demonstrated 95% family–wise confidence level of the results. The plotted homogeneity of variances of the results is shown in Figure 11.

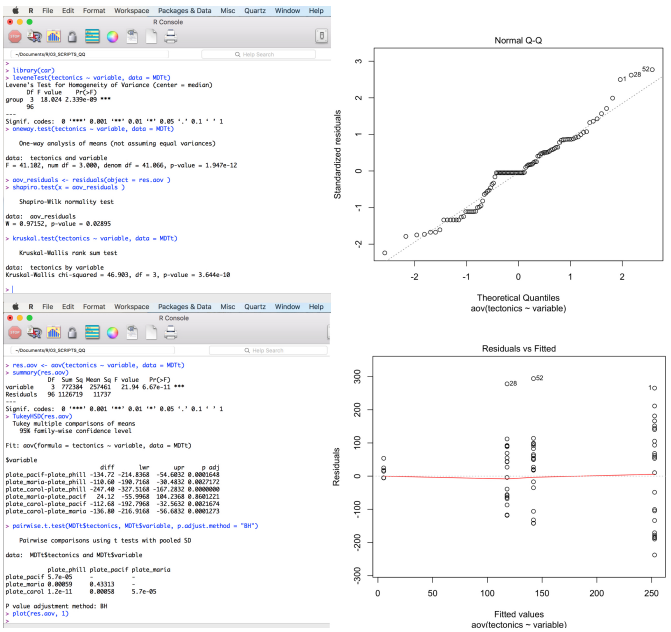

Figure 11. ANOVA hypothesis testing

Results

As can be drawn from the Figure 5, Mariana plate has the highest Kernel density of depth distribution values, followed by the Philippine plate, then Pacific and Caroline, respectively. Kernel density distribution has been shown in a combined plot enabling to compare the overlapping and maximal aptitude of the density curves for both the profiles and depth data distribution by four tectonic plates. The major trend of the trench angles located on the Pacific plate has downward general line trend. The Philippine tectonic plate, on the contrary, has a minimal peak by profiles #14–21, and then moving upwards. The highest value for Caroline plate has profile #23, while the maximal level for Mariana plate has profile #7. The Philippine tectonic plate, on the contrary, has a minimal peak by profiles #14–21, and then moving upwards. The highest value for Caroline plate has profile #23, while the maximal level for Mariana plate has profile #7.

From the composition charts (Figure 8), drawn to compare the slope angles and aspect degree by bathymetric profiles, one can see the unique patters of these categories for the trench. The basic concepts of

the close interrelation of various environmental factors (geological structure, bathymetric patters, tectonic plates location and subduction, spreading, transform fault, depths) affecting Mariana Trench formation is reflected in this paper. The technical tools consist of combination of R programming, GIS spatial analysis and mathematical algorithms for data processing and modelling. A comparison of the properties of the oceanic crust and ophiolite sections, sedimentary thickness and depth distribution across bathymetric profiles enabled to make a better understanding of the ties among geospatial factors influencing Mariana Trench formation. In turn, it facilitate a more reasonable forecast of the possible position of the mineral deposits fossils within Mariana Trench, for instance in the Philippine Sea tectonic plate.

The algorithm based objective analysis of the Mariana Trench bathymetric unevenness enables to better understand its morphology and structure. The special section of this research is focused on the detailed technical evaluation of the R and LaTeX scripts and description of their functionality. The thorough description of the workflow steps is accompanied by the written most important R scripts and codes provided in full, what makes this work repeatable. The discussion of the advantages of the methods enables to have a representative information on the key procedures of the research. As demonstrated in this paper, the combination of the statistical analysis by R programming language, modelling algorithms and GIS geospatial analysis provide the most effective methods for studying ocean floor and hadal trenches, the least reachable research objects on the Earth.

Discussion

The study of the structure of the Mariana Trench has a great scientific and practical interest. First, it relates to the testing theory of plate tectonics. Second, the estimation of the time and place of where earthquakes and tsunami may arise is of great importance for the environmental risk reduction, natural hazard protection and prevention. Moreover, the distribution of the submarine volcanoes is important, as they are one of the principal dynamic forces acting on the Earth's crust. Finally, ocean mining of the deep–seabed mineral resources has enormous economic potential.

The aim of the presented research was to identify bathymetric cross section profiles of the Mariana Trench that show similarity by their complex parameters, so that they could be grouped into patters. Using combination of R programming, LaTeX visualization and GIS, supported by the traditional geologic research, this study enabled to identify variabilities among factors affecting Mariana trench morphology. The accuracy of the performed statistical measurements is impacted by the chosen methodology and the input data: geomorphological maps, geological attribute tables, bathymetric maps, other geospatial thematic layers.

The question of the formation and structure of the ocean trenches has long attracted oceanographers. A variety of research work has been performed applied for various aspects of the hadal trenches: to measure trench depth, to assess the volumes of possible hidden resources in the abyssal depths, to analyse pelagic and biotic communities, to predict earthquakes frequencies. Nowadays, sound statistical methods are elaborated as important additions for traditional GIS methods of geospatial analysis of the deep ocean trenches. Without machine learning algorithms, a geospatial analysis of the big data cannot be considered as accurate and may lead to minor or major shortcomings and errors. Machine learning is therefore indispensable for such remotely located areas as abyssal depths.

The principle of R based geospatial data mining for the ocean research is to perform fusion and integration of multi–source data that allows integration of the heterogeneous information from a variety of sources for the possibility of in–depth analysis. In such a way, a multi–dimensional data processing is possible including following thematic layers: tectonic (continental margin plates), geological (geologic structure), lithological (sedimentation) and oceanological (direction of the deep currents). Automatic reflection of the multi–variant data by the machine learning methods enables to speed up our understanding of the oceans, as well as to increase the accuracy of the research.

Conclusion

The paper presents an application of the R language towards geological studies with a specific case of marine geological features, such as Mariana Trench. The brief sketch on the structure of the main environmental factors of the Mariana Trench has been presented in introduction section, briefly summarizing tectonic plates' movements, geological structure, sediment conditions, bathymetry, lithological properties and other features of the Mariana Trench system. It furthermore reviews the impact of the environmental determinants on the Mariana Trench formation and structure, for instance, the relationship of the transform faults with rift valleys, location of the large igneous polygons. The methodology chapter is focused on the technical approaches of the study of the trench morphology by means of R programming, statistical analysis and GIS. The techniques of the classification are presented, including key R programming scripts. The results chapter highlights the findings received by the application of the technical methods (R programming, statistical algorithms and GIS) with geological analysis of the Mariana Trench seafloor. The graphical and cartographic materials presented in this research support results and findings. Current work demonstrates that the combination of R language, statistical algorithms and Quantum GIS is a highly effective decision for geospatial research aimed at processing multi–dimensional geospatial data sets.

As a recommendation for further research focused on ocean trenches, a traditional methods of the seafloor studies at the GIS level should always include a set of statistical analysis tools such as R, or alternatively, python (e.g. ScyPy, NumPy, matplotlib), Matlab, Octave. Studies of the spatial data series are usually accompanied by the graphics in view of the facetted plots that can facilitate comparison of the data by groups. This is especially useful while collecting multi–dimensional various geophysical data. Since there are many deep interconnections among factors affecting the formation, morphology and development of the trench, mathematical algorithms and objective analyses based approaches are to be used to study marine geo–systems.

Conflict of Interest

The author declares that there is no conflict of interest.

Acknowledgements

The study has been funded by the China Scholarship Council (CSC), State Oceanic Administration (SOA), Marine Scholarship of China [Grant # 2016SOA002, 2016].

References

Butuzova, G. Y. (2003). Hydrothermal–Sedimentary Ore Formation in the World Ocean. *Geos*, Moscow, Russia. 136p.

Dubinin, E. P. & Ushakov, S. A. (2001). Oceanic Rift Genesis. *Geos*, Moscow, Russia. 293p.

Gardner, J. V., Armstrong, A. A., Calder, B. R. & Beaudoin, J. (2014). So, How Deep Is The Mariana Trench? *Marine Geodesy*, ***37***(1): 1–13.

Garfunkel, Z., Anderson, C. A. & Schubert, G. (1986). Mantle Circulation and the Lateral Migration of Subducted Slabs. *Journal of Geophysical Research*, ***91***: 7205–7223.

Gurevich, E. G. (1998). Metalliferous Sediments in the World Ocean. Nauchnyy Mir, Moscow, Russia. 340p.

Horleston, A. C. & Helffrich, G. R. (2012). Constraining Sediment Subduction: A Converted Phase Study of the Aleutians and Marianas. *Earth and Planetary Science Letters*, ***359–360***: 141–151.

Husson, L. (2012). Trench Migration and Upper Plate Strain over a Convecting Mantle. *Physics of the Earth and Planetary Interiors*, ***212–213***: 32–43.

Ishibashi, J., Tsunogai, U., Toki, T., Ebina, N., Gamo, T., Sano, Y., Masuda, H. & Chiba, H. (2015). Chemical Composition of Hydrothermal Fluids in the Central and Southern Mariana Trough Backarc Basin. *Deep–Sea Research Part II: Topical Studies in Oceanography*, ***121***: 126–136.

Lin, J. W. B. (2008). Qtcm 0.1.2: A Python Implementation of the Neelin–Zeng Quasi–Equilibrium Tropical Circulation Model. *Geoscientific Model Development*, ***1***: 315–344.

Marta–Almeida, M., Ruiz–Villarreal, M., Otero, P., Cobas, P., Peliz, A., Nolasco, R., Cirano, M. & Pereira, J. (2011). OOF3: A Python Engine for Automating Regional and Coastal Ocean Forecasts. *Environmental Modelling & Software*, ***26***: 680–682.

Oliphant, T. E. (2007). Python for Scientific Computing. *Computing in Science & Engineering*, ***9***: 10–20.

R Core Team (2018). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. https://www.R–project.org

Roberts, J. J., Best, B. D., Dunn, D. C., Treml, E. A. & Halpin, P. N. (2010). Marine Geo–Spatial Ecology Tools: An Integrated Framework for Ecological Geoprocessing with ArcGIS, Python, R, MATLAB, and C++. *Environmental Modelling & Software*, ***25***: 1197–1207.

Uyeda, S. & Kanamori, H. (1979). Back–Arc Opening and the Mode of Subduction. *Journal of Geophysical Research*, ***84***: 2017–2037.

Warner, J. C., Perlin, N. & Skyllingstad, E. D. (2008). Using the Model Coupling Toolkit to Couple Earth System Models. *Environmental Modelling & Software*, ***23***: 1240–1249.