IJIAM
International Journal of Informatics
and Applied Mathematics

International Journal of Informatics and Applied Mathematics
e-ISSN:2667-6990 Vol. 2, No. 1, 1-26

K-means Clustering in R Libraries {cluster} and {factoextra} for Grouping Oceanographic Data[⋆]

Polina LEMENKOVA[1]

Ocean University of China, College of Marine Geo-sciences, Qingdao, 238 Songling
Rd., Laoshan, 266100, Shandong Province, Peoples Republic of China
pauline.lemenkova@gmail.com

Abstract. Cluster analysis by k-means algorithm by R programming applied for the geological data analysis is the scope of the presented paper. The research object is the Mariana Trench, a hadal trench located in west Pacific Ocean. The study evaluates the similarity of the geological data by the analysis of their attributes. The original observation data set contained samples varying in parameters: geology (sediment thickness), tectonics (locations on the tectonic plates), volcanism (igneous volcanic areas), bathymetry (depth ranges) and geomorphology (slope steepness and aspect). The data pool was divided to the clusters using k-means algorithm with aim to detect similarities. Clustering was chosen as a main statistical method, since it enables detecting similar groups within the original data set by unsupervised classification. Technically, the research was performed using R language and its statistical libraries. The main R libraries include {cluster}, {factoextra}; minor libraries include {ggplot2}, {FactoMiner}, {openxlsx}, {carData}, {rio}, {car} and {flashClust}. Several clusters were tested from two to seven, the optimal number is defined as five. The results show visualized computations:

[⋆] Supported by China Scholarship Council (CSC), State Oceanic Administration (SOA), Marine Scholarship of China, Grant Nr. 2016SOA002, Beijing, P.R.C.

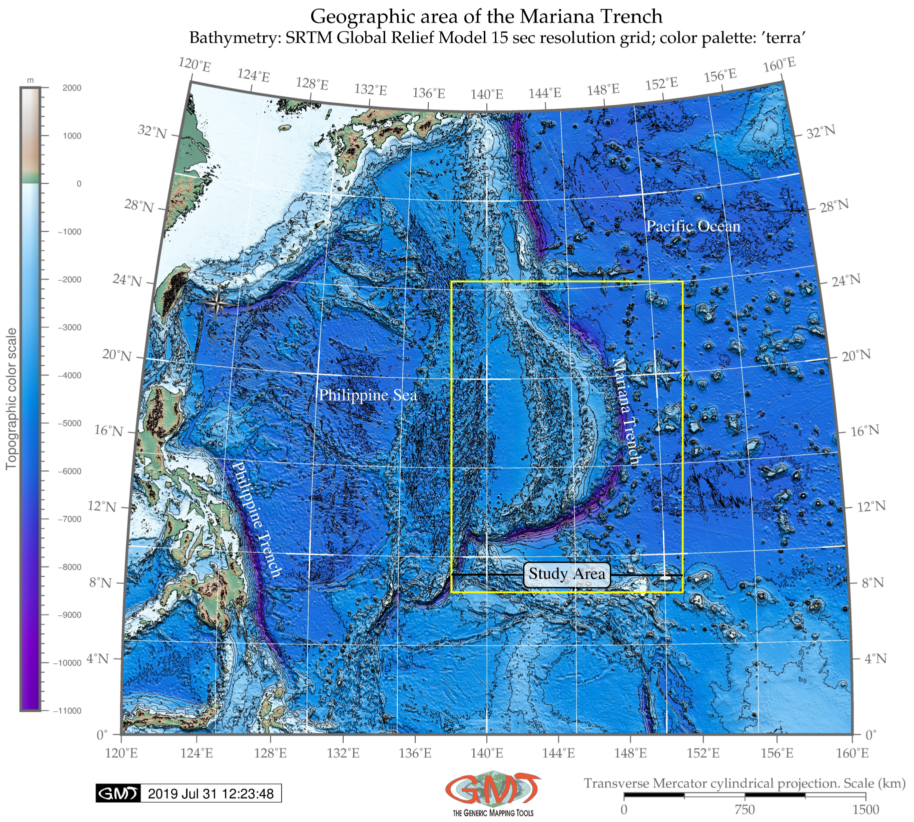
correlation matrix of the factors; comparison of the *bi*-factors showing pairwise correlation; pairwise comparative analysis showing influence of the variables as *bi*-factors: sediment thickness correlating with slope angles; correlation of the volcanic igneous areas with slope angles and aspect degree. Four variables affect geomorphology: slope angle, sediment thickness, aspect degree, bathymetry and volcanism. The paper includes listings of R programming codes for repeatability of the algorithms in similar research.

Keywords: R · programming language · statistics · geospatial data · k-means clustering · cluster analysis · data grouping · marine geology

1 Introduction

Machine learning techniques by R is widely used for the statistical data analysis in various domain: Information Technologies (IT), economics, finance, programming, data sciences. However, their application in the Earth sciences is less popular comparing to the traditional GIS approach. Existing application focus on the structural geology and general aspects of the geosciences [42,34].

Current paper aims to contribute towards methodological development of the statistical data analysis in marine geology by presenting clustering method for data analysis with an example of the *k*-means clustering. Clustering as a statistical algorithm applied for the detection of the similarity within the data set. The data were divided into groups for analysis and modelling using clustering techniques explained in details below. Initial geospatial data analysis and mapping was based on the Quantum GIS (QGIS), Generic Mapping Tools (GMT). Combination of GIS with R programming enabled effective modelling of the data set visualized the abstraction of the real phenomena of the Earth. This provides insights to the underlying geological processes aimed at interpretation and analysis of the hidden processes, such as correlation of the environmental parameters with geomorphic patterns of the deepest regions of the Earth. The R language has significant number of the statistical libraries designed for machine learning. Using statistical algorithms of R in the geosciences increases the precision of the data modelling and exploring. The particular advantage of the using statistical algorithms specifically for marine geological consists in the fact that it can highlight underlying patterns and phenomenal relationships and correlations in the data sets that are difficult to find otherwise, due to the specific nature of the study object. In fact, the deep-sea trenches are the least reachable geological objects on the planet that can only be studied by the remote sensing techniques or using machine learning and data modelling. Therefore, R libraries were tested for the for statistical data analysis in this research. Specifically, the data modelling by {cluster} and {factoextra} packages used for the cluster analysis was demonstrated in this paper.



Geographic area of the Mariana Trench

Bathymetry: SRTM Global Relief Model 15 sec resolution grid; color palette: 'terra'

Fig. 1. Study area: Mariana Trench, located in the west Pacific Ocean, eastward from the Philippine Sea and southward from Japan. Mapping: GMT

2 Study area: brief description

A specific case study of this research is Mariana Trench, located in west Pacific Ocean, as illustrated on Fig. 1 mapped using GMT by Shuttle Radar Topography Mission (SRTM). The study goal is to analyze the variability of the geomorphic shape structure by means of the data analysis. The importance of the application of the advanced computing methods for the analysis of the submarine geomorphology and variations of the depths in the ocean trenches is caused by the inaccessibility of the study object. With depths reaching maximal values of 11,000 meters [41], Mariana Trench is the least accessible geologic phenomena on the Earth. However, using data sets derived from the electronic maps we can analysis the structure of the trench that creates conditions for the submarine ecosystems where life exists in the deepest hadal zones.

The existence of the organisms at such extreme depths strongly depends on the variety of the factors: geomorphology, geological substrate, turbulence and motion of the ocean streams upbringing nutrients for the living organisms. In turn, these are strongly influenced by the physical oceanographic settings and geomorphic structure of the trench [35]. Hadal fauna survive on the fringes of the two extremes caused by the hydrostatic pressure and remoteness from the surface-derived food supply [18]. The interconnectivity of the environmental factors affecting hadal ecosystems is described by [49], who reviewed aspects of the hadal biogeochemistry: the effect of food supply on the hadal ecosystems, carbon cycle in the sub-seafloor under high hydrostatic pressure and pollution in the trenches. Consequently, the communities of the trenches representing spatially isolated environments highly distinct from the shallower areas.

High sedimentation rates and biomass of the trenches, intense microbial activities and chemosynthetic communities play crucial role in the global ocean biogeochemical cycles.Other environmental characteristics of the Mariana Trench include a nearly uniform distributional pattern of heterotrophic bacteria in the trench interior [40]. As briefly demonstrated above, factors affecting such complex structure of the submarine ecosystem as Mariana Trench are highly diverse and interconnected. Analysis of these phenomena becomes possible by the application of the numerical modelling, advanced methods of the statistical data analysis and machine learning.

3 Materials and Methods

The methodological scheme includes several steps of the statistical analysis of the geomorphology of the Mariana Trench:

- Mapping study area using GMT based on the SRTM raster grid (Fig. 1);
- QGIS based processing of the geospatial data; digitizing cross-section profiles across the trench;
- Creating initial data set: converting coordinates to the Universal Transverse Mercator (UTM); reading depths of the sample points into the table;

- Converting initial table to the R environment as Data Frame (DF); The Non available (numbers) (NA) values were removed from the original data frame;
- Statistical analysis of data distribution: calculation and visualization of the data distribution by the bathymetric profiles and geological settings;
- Clustering by k centers in clustering algorithm (k)-means method aimed at data partition, grouping and sorting;
- Plotting Principal Component Analysis (PCA) for the bathymetric data.

The initial geospatial analysis of the geological data was performed using QGIS software and GMT scripting toolset based on the available manual and technical literature [48,47]. The visualized and processed data are based on the available geographical vector layers [46] and raster SRTM [12]. The geospatial analysis aimed to extract regular impact factors of the geomorphology, whereas computing and visualizing cluster analysis extracts the groups and classes of data from a total pool of the observation samples. Statistical analysis has been performed by means of R [33].

3.1 Analysis of Data Distribution

The data analysis has been performed using existing methods of the statistical analysis [34,3,38,24,23,21]. Data exploration is focused on the analysis of the sampled data set of the Mariana Trench. The first research step included plotting and visualizing data distribution aimed to understand how variables interact with each other, to show data distribution and detect outliers. To this end, the descriptive statistics on the data was computed using {ggboxplot} library of R for the notched box plots and {ggplot} library for the histograms.

Notched Box Plots The notched boxplot (Fig. 2) showing distribution of the depths with majority of the data from -3000 to -5000 m and outliers (the deepest and shallowest points) show that generally, the depth increase from the profile 1 to 16 the with the deepest values at profiles 10 and 11. The original data frame containing depths of the trench (Marina trench depth values data frame (MDepths)) was processed using R function 'ggboxplot'.

```
  #Part 1
  #step-1. generating dataframe from the raw table Depths.csv
MDepths <- read.csv("Depths.csv", header=TRUE, sep = ",")
  # step-2. Cleaning dataframe from the NA values
MDepths_df <- na.omit(MDepths)
row.has.na <- apply(MDepths_df, 1, function(x){any(is.na(x))}) # check up the NA
sum(row.has.na) # sup up theNA, result: [1] 0
head(MDepths_df) # look up dataframe
  # Part 2: generating whisker boxplot using dataframe MDepths_df.
  # step-3. Adding palette, lines type, Chinese fonts.
p<- ggboxplot(
  MDepths_df, title="Mariana Trench, Profiles 1-25.",
  subtitle = "Notched Boxplot for Data Groups by 25 Profiles with Outliers)",
  caption = "Statistics Processing and Graphs: \nR Programming. Data Source: \ac{QGIS}",
  x = "profiles", y = "depths", width = 0.8, notch = TRUE,
  fill = "profiles", linetype = 1, size = .1, outlier.colour = "grey44",
  palette = c("magma"), orientation = "horizontal")
# step-4. Adding theme and palette using default format
```

```
boxplot_Mariana<- p + theme()
```

Listing 1.1. R code for notched box plots

As can be seen from the notched boxplot (Fig. 2), the shallowest parts of the trench are located in the profiles 24 and 25 on the south, where Mariana Trench crosses Yap Trench, the oceanic trench near the Yap Island in west Pacific Ocean. The geomorphological analysis shows the depths records where each value contribute to the observations data pool of the bathymetric patterns.

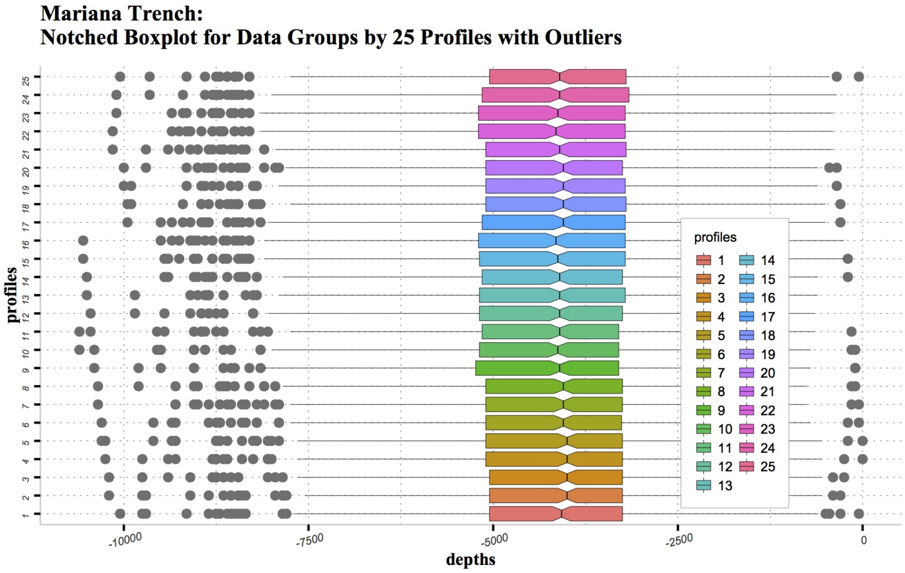


Fig. 2. Analysis of Data Distribution: Notched Box Plots. Plotting: R.

For instance, the gradual decrease of the values can be noticed for the profiles Nr. 1 to 16 which indicates the decrease of the depths in southward direction. The outliers seen on the profiles as gray circles show depth values that lie in a data set on the extreme values. Such values indicate places where the profiles crossed the islands or the deepest parts of the trench and are not normal for the profile shape. The outliers were removed from the data set on the next step of data analysis, because they do not show natural geomorphic shape of the trench. After the outliers were discarded from the data series the segmentation analysis was performed and NA values were ignored.

Histograms The histogram plots (Fig. 3) show frequency distribution of the depth values by the observations recorded in the profiles. Each histogram shows distribution of how often each depth record appear in a bathymetric data set.

```
MDepths <- read.csv("Depths.csv", header=TRUE)
X01<- MDepths[,01]
X01<-X01[!is.na(X01)]
as.data.frame(X01)
dat01<- data.frame(X01)
p01<-ggplot(dat01, aes(X01)) +
  labs(title = "Profile Nr.01", x = "Depths, m", y = "Density") +
  theme() +
    scale_x_continuous(breaks = pretty(dat01$X01, n = 4), minor_breaks = seq(min(dat01$X01),
          max(dat01$X01), by = 500)) + scale_y_continuous(breaks = scales::pretty_breaks(n =
          4),labels = scales :: percent) +
    scale_fill_distiller(palette = "RdGy") +
    scale_color_manual(name = "Statistics:", values = c(median = "purple", mean = "green4",
          density = "blue", norm_dist = "black")) +
  geom_histogram(binwidth = 200,aes(fill = ..density..,x = dat01$X01,y = ..density..),color =
         "blue",size = .1) +
  stat_function(fun = dnorm, args = list(mean = mean(dat01$X01), sd = sd(dat01$X01)), lwd =
        0.2, color = ’black’) +
  stat_density(
      geom = "line", size = .3, aes(color = "density")) +
      geom_vline(aes(color = "mean", xintercept = mean(X01)), lty = 4, size = .3) +
      geom_vline(aes(color = "median", xintercept = median(X01)), lty = 2, size = .3) +
      geom_vline(aes(color = "norm_dist", xintercept = dnorm(X01)), lty = 2, size = .3)
```

Listing 1.2. R code for histograms plotting. Step-1: plotting 1 profile

Therefore, the data pool included all normal observations to show natural changes in the ocean seafloor depths.

```
library(cowplot)
figure <-plot_grid(
p01 + theme(legend.position="none"),
p02 + theme(legend.position="none"),
# continue sequently until profile Nr. 25:
p25 + theme(legend.position="none"),
labels = c("1", "2", "3", "4", "5", "6", "7", "8", "9", "10", "11", "12", "13", "14", "15",
"16", "17", "18", "19", "20", "21", "22", "23", "24", "25"),
ncol = 4, nrow = 7)
```

Listing 1.3. R code for combining 25 individual histogram plots on one facetted plot

The upper limit of the notched box plot (Fig. 2) shows the shallowest values of the depths located in the proximity of the Mariana arc islands, and the lowest values show the deepest values recorded in the south-western part of the trench. The technique of the R code used for generating notched box plot is used from the 'ggplot' documentation: ggboxplot. The InterQuartile Range (IQR) might be added as additional parameter using 'median_iqr', but in the scope of this research was omitted as not required. The R code used for plotting Fig. 2 is shown in the Listing 1.1. The annotation for the histogram figure were added using R Listing 1.4.

```
figure_all_cowplot<- annotate_figure(figure,
    top = text_grob("Mariana Trench Bathymetry: Histograms of Depth Distribution", color = "
          lightsteelblue4", face = "bold", size = 10),
    bottom = text_grob("Data processing: \n R, QGIS", color = "blue", hjust = 1, x = 1, face
          = "italic", size = 8),
    left = text_grob("Figure arranged using R, ggpubr", color = "slategray4", size = 8, rot =
           90),
    right = text_grob("1000-km length profiles", color = "slategray4", size = 8, rot = 270),
    fig.lab = "Profiles 1-25", fig.lab.face = "bold", fig.lab.size = 8, fig.lab.pos = "bottom
          .left")
```

```
ggsave("figure_all_cowplot.pdf", device = cairo_pdf, fallback_resolution = 300, width = 210,
     height = 297, units = "mm")
```

Listing 1.4. R code for adding annotations for the facetted histogram on Fig. 3

Because the data shows hadal trench, the most frequent values (those on the peaks of the histograms) are ranged between the -3,000 and -6,000 m. Shallow values have low frequency only showing the samples where profiles cross the island arc. The deepest values (deeper than -7,000 m) are recorded in the south-western part of the trench. The numbering of the profiles goes consequently from the north (profile Nr. 1) to the south-west (profile Nr. 25) following the arc-shape form of the trench (see Fig. 1). The histograms show the outliers, skewness, median, mean, average, maximal, minimal and quartile distribution for the depth values in each of the 25 profiles. The deepest values are located in the profiles Nr. 21 and 22 where the deepest place of the Earth is detected: the Challenger Deep (-10,898 m). The majority of the profiles show Gaussian normal distribution (e.g., profiles Nr. 1, 5, 7-11, 15-20, 23-25). Some other profiles show bimodal distribution n (double-peaked), for instance, profiles Nr. 2-4, 12-14. Multimodal distribution (similar to plateau in its geometric shape) can be noted on the profiles Nr. 6 and 22.
Clearly skewed distribution, asymmetrical in shape is noticed for the profiles 21 and 22 (both right-sided), profile Nr. 24 (left-sided) The skewness is caused by the geomorphological shape of the trench, because Mariana Trench has crescent-like shape form (see Fig. 1) and the geological substrate of the rocks, together with other factors (submarine oceanic currents causing erosion, sedimentation processes, tectonic movements causing plates subduction, etc.) affects geomorphology. Profiles Nr. 16 and 25 demonstrate edge peak distribution with small peak at one 'tail' of the histogram showing increase in depths deeper than 8,000 and 7,000 mm, respectively.
For each of the 25 bathymetric profile various colours are taken for the following statistical data of depth values along the profile: black curves stands for normal distribution, 'blue' curves for density distribution. Vertical dashed lines represent purple: median values, green: mean values. The histograms have been drawn using R library {ggplot2}.

First, a single histogram for each of bathymetric 25 profiles was created. The R script used to plot a histogram is (here: for the profile Nr. 1, further applied for every one from 25 profiles by changing the name of the file from 01 and so on to 25) is shown in the Listing 1.2.
Then, using this code further 25 profiles were plotted, consequently, p01, p02, p25. On the next step the combination of the 25 profiles on one layout (Fig. 3) was done using R code by library {cowplot} shown in Listing 1.3.
Finally, generated plot of the 25 histograms (Fig. 3) illustrates the frequency of the depths in a data set. The histogram bins show variations in the samples.

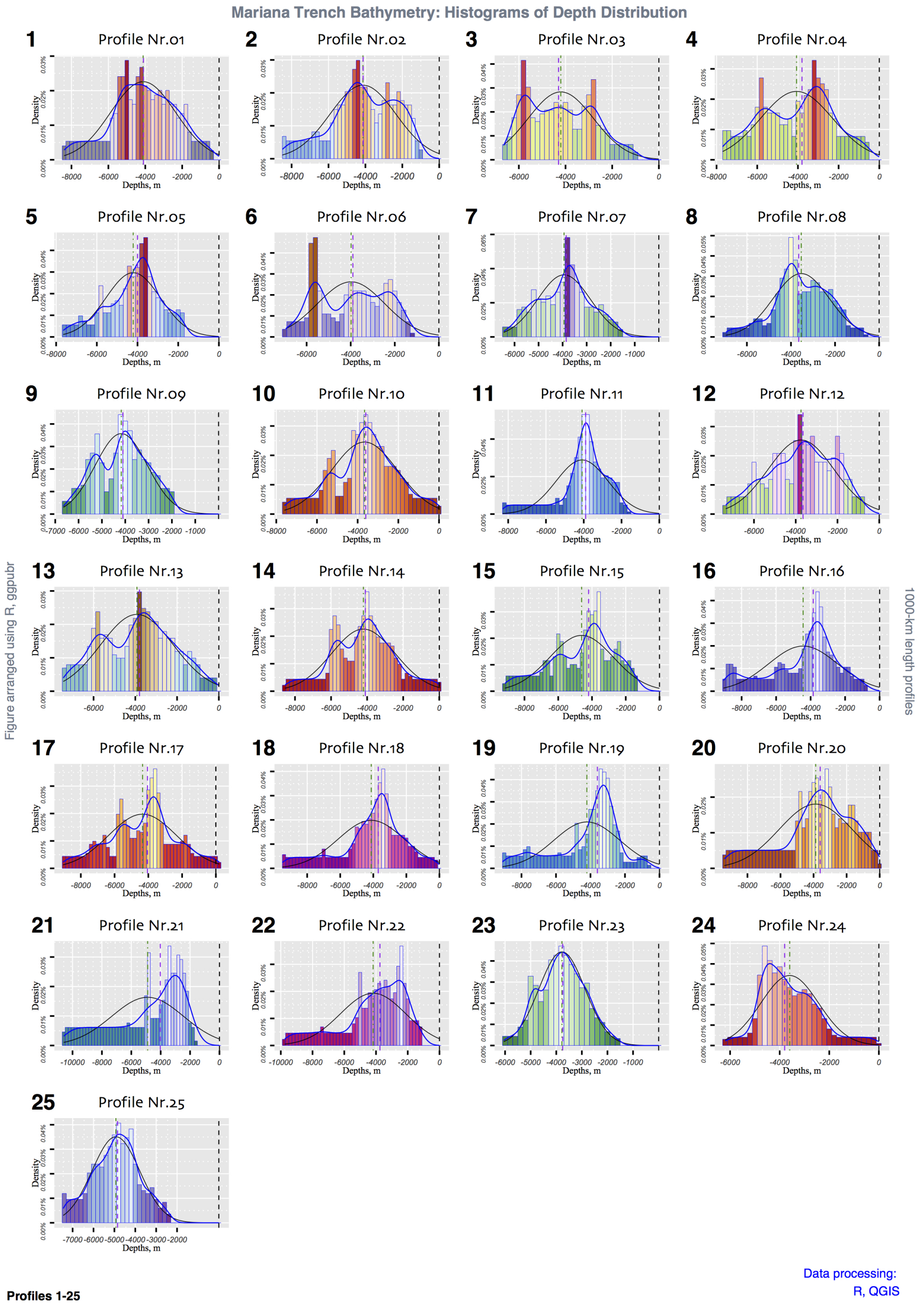


Fig. 3. Faceted plot of the histograms showing samples (depths) by the 25 cross-section profiles. X axis: depths, Y axis: frequency of the data distribution

3.2 Cluster Analysis by R Libraries {factoextra} and {cluster}

Using clustering technique is well documented in the statistical research and as as in geological applications [6,39,7,11,9,43,22,13,20,31,33]. Cluster analysis used in this work aimed at finding groups of the observation data that show geomorphic shapes of the 25 profiles located across four tectonic plates crossing Mariana Trench: Mariana, Philippine, Caroline and the Philippine Sea. To this end, cluster analysis was performed using {factoextra} and {cluster} libraries of R as visualized by the prinscreens on Fig. 4 and Fig. 5.

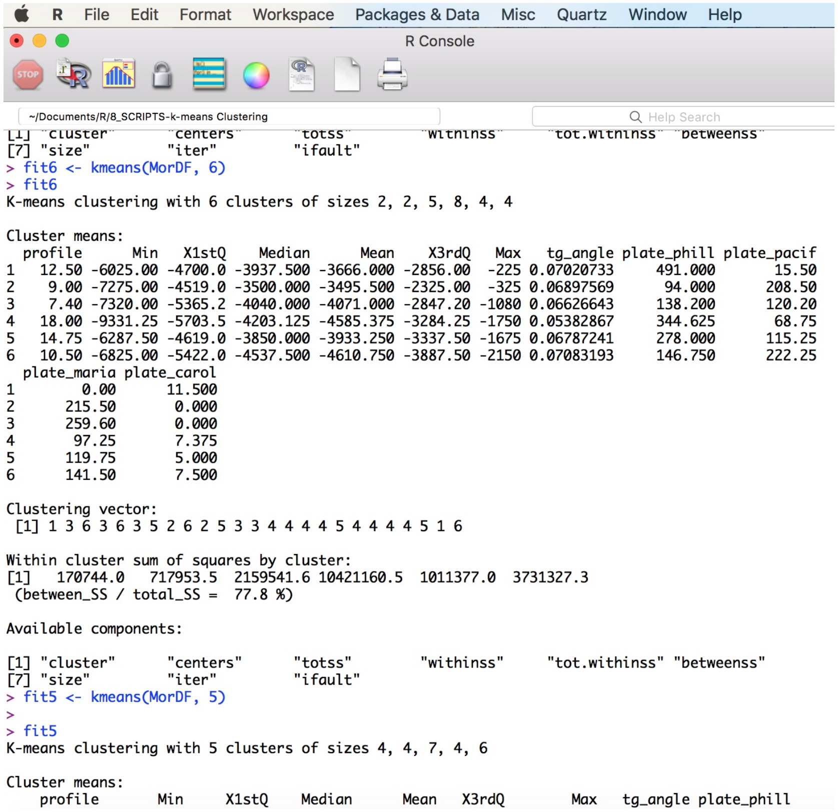
```
# PART 1: create data.frame with geomorphology data
  # step-1. Load table, create dataframe
MorDF <- read.csv("Morphology.csv", header=TRUE, sep = ",")
head(MorDF)
summary(MorDF)
# PART 2: Clustering
  #  step-2. Create several examples of cluster analysis with various numfig:9ber of cluster
        centers (k). Here: 5
k2MorDF <- kmeans(MorDF, centers = 5, nstart = 25)
str(k2MorDF)
k2MorDF
fviz_cluster(k2MorDF, data = MorDF)
  # step-3. Creaing objects for each of the plots (1 to 7, here: example for plot 2):
p2 <- fviz_cluster(k2MorDF, geom = "point", data = MorDF) + ggtitle("Nr. of centers k = 2") +
theme(plot.title = element\_text(size = 10), legend.title = element_text(size=8),
  legend.text = element\_text(colour="black", size = 8))
  # step-4. Combine all plots on one layout:
figure <-plot_grid(p2, p3, p4, p5, p6, p7, labels = c("1", "2", "3", "4", "5", "6"), ncol =
      2, nrow = 3)
  # step-5. Add legend, title and theme:
ClustersMariana6 <- figure +
  labs(title="Mariana Trench, Profiles Nr. 1-25.",
  subtitle = "Geomorphological Cluster Analysis (k-means)",
  caption = "Statistics Processing and Graphs: \nR Programming. Data Source: QGIS") +
  theme(plot.title = element_text(family = "Arial", face = "bold", size = 12),
plot.subtitle = element\_text(family = "Times New Roman", face = "bold", size = 10)).
```

Listing 1.5. R code for k-means cluster analysis

In this process, the algorithms is based on the dividing data set according to their similarity into observation subsets, or clusters, where samples are similar to those in the same cluster they belong to, but differ from those in other clusters: Fig. 6. The circles in the top left of the four subplot figures (Fig. 6) are clearly closer to each other while being far away from the others. The same is true for the polygons visualized on Fig. 7.

k-means Clustering Method The k-means clustering is a machine learning technique that is developed for the data partition into defined Number (Nr) of clusters [10,1]. In the current research clustering has been done using R libraries {factoextra} and {cluster}.

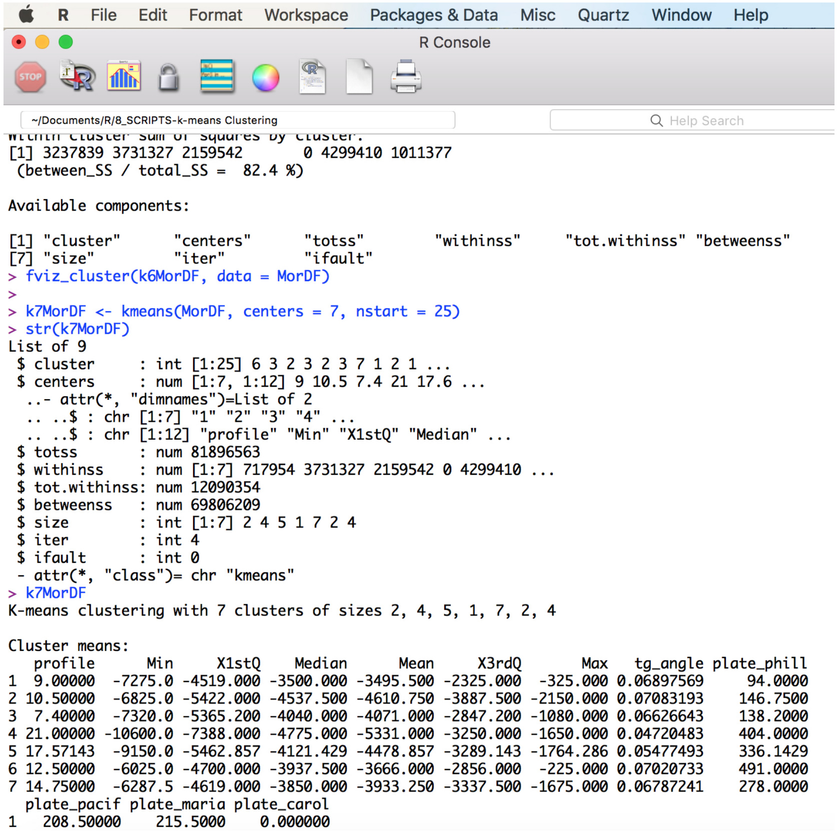
Conceptual aim of the procedure The goal of the k-means clustering was to perform partition of the dataset of the observations across Mariana Trench into clusters, in which each bathymetric observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. The performed clustering is sensitive to the initial random selection of the cluster centers. Therefore, several



```
[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> fit6 <- kmeans(MorDF, 6)
> fit6
K-means clustering with 6 clusters of sizes 2, 2, 5, 8, 4, 4

Cluster means:
  profile      Min   X1stQ    Median      Mean    X3rdQ   Max   tg_angle plate_phill plate_pacif
1   12.50 -6025.00 -4700.0 -3937.500 -3666.000 -2856.00  -225 0.07020733     491.000       15.50
2    9.00 -7275.00 -4519.0 -3500.000 -3495.500 -2325.00  -325 0.06897569      94.000      208.50
3    7.40 -7320.00 -5365.2 -4040.000 -4071.000 -2847.20 -1080 0.06626643     138.200      120.20
4   18.00 -9331.25 -5703.5 -4203.125 -4585.375 -3284.25 -1750 0.05382867     344.625       68.75
5   14.75 -6287.50 -4619.0 -3850.000 -3933.250 -3337.50 -1675 0.06787241     278.000      115.25
6   10.50 -6825.00 -5422.0 -4537.500 -4610.750 -3887.50 -2150 0.07083193     146.750      222.25
  plate_maria plate_carol
1        0.00      11.500
2      215.50       0.000
3      259.60       0.000
4       97.25       7.375
5      119.75       5.000
6      141.50       7.500

Clustering vector:
 [1] 1 3 6 3 6 3 5 2 6 2 5 3 3 4 4 4 4 5 4 4 4 4 5 1 6

Within cluster sum of squares by cluster:
[1]   170744.0   717953.5  2159541.6 10421160.5  1011377.0  3731327.3
 (between_SS / total_SS =  77.8 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> fit5 <- kmeans(MorDF, 5)
>
> fit5
K-means clustering with 5 clusters of sizes 4, 4, 7, 4, 6

Cluster means:
    profile      Min     X1stQ    Median      Mean   X3rdQ        Max   tg_angle plate_phill
```

Fig. 4. Printscreen of the *k*-means clustering process in R: fitting data set

clusters were tested from two to seven. This function provides a solution using a hybrid approach by combining the hierarchical clustering and the *k*-means methods. After the trial testing was done, a kmeans++ algorithm [2] was additionally used as an improved version of the *k*-means clustering enabling to choose the initial values for the *k*-means clustering algorithm in a more optimal way.



```
Within cluster sum of squares by cluster:
[1] 3237839 3731327 2159542       0 4299410 1011377
 (between_SS / total_SS =  82.4 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"     "tot.withinss" "betweenss"
[7] "size"         "iter"         "ifault"
> fviz_cluster(k6MorDF, data = MorDF)
>
> k7MorDF <- kmeans(MorDF, centers = 7, nstart = 25)
> str(k7MorDF)
List of 9
 $ cluster     : int [1:25] 6 3 2 3 2 3 7 1 2 1 ...
 $ centers     : num [1:7, 1:12] 9 10.5 7.4 21 17.6 ...
  ..- attr(*, "dimnames")=List of 2
  .. ..$ : chr [1:7] "1" "2" "3" "4" ...
  .. ..$ : chr [1:12] "profile" "Min" "X1stQ" "Median" ...
 $ totss       : num 81896563
 $ withinss    : num [1:7] 717954 3731327 2159542 0 4299410 ...
 $ tot.withinss: num 12090354
 $ betweenss   : num 69806209
 $ size        : int [1:7] 2 4 5 1 7 2 4
 $ iter        : int 4
 $ ifault      : int 0
 - attr(*, "class")= chr "kmeans"
> k7MorDF
K-means clustering with 7 clusters of sizes 2, 4, 5, 1, 7, 2, 4

Cluster means:
   profile      Min     X1stQ    Median      Mean     X3rdQ       Max   tg_angle plate_phill
1  9.00000  -7275.0 -4519.000 -3500.000 -3495.500 -2325.000  -325.000 0.06897569     94.0000
2 10.50000  -6825.0 -5422.000 -4537.500 -4610.750 -3887.500 -2150.000 0.07083193    146.7500
3  7.40000  -7320.0 -5365.200 -4040.000 -4071.000 -2847.200 -1080.000 0.06626643    138.2000
4 21.00000 -10600.0 -7388.000 -4775.000 -5331.000 -3250.000 -1650.000 0.04720483    404.0000
5 17.57143  -9150.0 -5462.857 -4121.429 -4478.857 -3289.143 -1764.286 0.05477493    336.1429
6 12.50000  -6025.0 -4700.000 -3937.500 -3666.000 -2856.000  -225.000 0.07020733    491.0000
7 14.75000  -6287.5 -4619.000 -3850.000 -3933.250 -3337.500 -1675.000 0.06787241    278.0000
  plate_pacif plate_maria plate_carol
1   208.50000    215.5000    0.000000
```

Fig. 5. Printscreen of the k-means clustering in R: data analysis by machine learning

The kmeans++ is an approximation algorithm for the NP-hard *k*-means problem enabling to avoid poor clusterings in the standard *k*-means algorithm.

The workflow procedure The workflow procedure include following steps:

- Computed clustering and fitting data in the k-clusters (Fig. 4 and Fig. 5);
- Computed centers (2 to 7 were tested, in total 6 centers) of each cluster (Fig. 6 and Fig. 7);
- Computed and visualized correlation matrix using criterion Baysiean Information Criterion (BIC) (Fig. 8 and Fig. 9);

– Pairwise standard scatterplot of k-means cluster correlation (Fig. 10);
– Plotted PCA (Fig. 11).

The k-means cluster analysis has been done using code provided in R Listing 1.5 applied for the Morphology data frame (MorDF).

Algorithm of the k-means Clustering by R The algorithm worked iteratively to assign each profile to one of the groups (2 to 7).

```
# Part 1.
k6MorDF <- kmeans(MorDF, centers = 6, nstart = 25)
str(k6MorDF)
k6MorDF
fviz_cluster(k6MorDF, data = MorDF)
# Part 2. Standard pairwise scatter plots to illustrate clusters compared to the original
     variables.
# Here:compare slope angles of the Mariana Trench on 4 tectonic plates.
  # 2.1.for theMariana Plate
PairM <- MorDF %>%
  as_tibble() %>%
  mutate(cluster = k6MorDF$cluster,
         profile = row.names(MorDF)) %>%
  ggplot(aes(x = plate_maria, y = tg_angle, color = factor(cluster), label = profile)) +
  geom_text()
  # 2.2.for the Philippine Plate
PairPh <- MorDF %>%
  as_tibble() %>%
  mutate(cluster = k6MorDF$cluster,
         profile = row.names(MorDF)) %>%
  ggplot(aes(x = plate_phill, y = tg_angle, color = factor(cluster), label = profile)) +
  geom_text()
PairPh
    # 2.3.for the Pacific Plate
PairPc<- MorDF %>%
  as_tibble() %>%
  mutate(cluster = k6MorDF$cluster,
         profile = row.names(MorDF)) %>%
  ggplot(aes(x = plate_pacif, y = tg_angle, color = factor(cluster), label = profile)) +
  geom_text()
PairPc
    # 2.4.for the Caroline Plate
PairC<- MorDF %>%
  as_tibble() %>%
  mutate(cluster = k6MorDF$cluster,
         profile = row.names(MorDF)) %>%
  ggplot(aes(x = plate_carol, y = tg_angle, color = factor(cluster), label = profile)) +
  geom_text()
    # PART-3. label each plot
p1<- PairM + ggtitle("MARIANA Plate; Trench Profiles 1:25; Trench Angles (tg(A/H)") + theme(
     plot.title = element_text(size = 8), legend.title = element_text(size=8), legend.text =
      element_text(colour="black", size = 8), axis.title = element_text(size = 8))
p2<- PairPh + ggtitle("PHILIPPINE Plate; Trench Profiles 1:25; Trench Angles (tg(A/H)") +
     theme(plot.title = element_text(size = 8), legend.title = element_text(size=8), legend.
     text = element_text(colour="black", size = 8), axis.title = element_text(size = 8))
p3<- PairPc + ggtitle("PACIFIC Plate; Trench Profiles 1:25; Trench Angles (tg(A/H)") + theme(
     plot.title = element_text(size = 8), legend.title = element_text(size=8), legend.text =
      element_text(colour="black", size = 8), axis.title = element_text(size = 8))
p4<- PairC + ggtitle("CAROLINE Plate; Trench Profiles 1:25; Trench Angles (tg(A/H)") + theme(
     plot.title = element_text(size = 8), legend.title = element_text(size=8), legend.text =
      element_text(colour="black", size = 8), axis.title = element_text(size = 8))
  # PART-4. combine 4 plats together
Pair_figure <-plot_grid(p1, p2, p3, p4, labels = c("1", "2", "3", "4"), ncol = 2, nrow = 2)
```

Listing 1.6. R code for the pairwise clusters comparison

The assignment is based on the morphometric features of the Mariana Trench across these profiles that were clustered based on their geomorphic feature similarity. The *k*-means cluster analysis was performed using set of the cluster centers as the initial cluster centers.

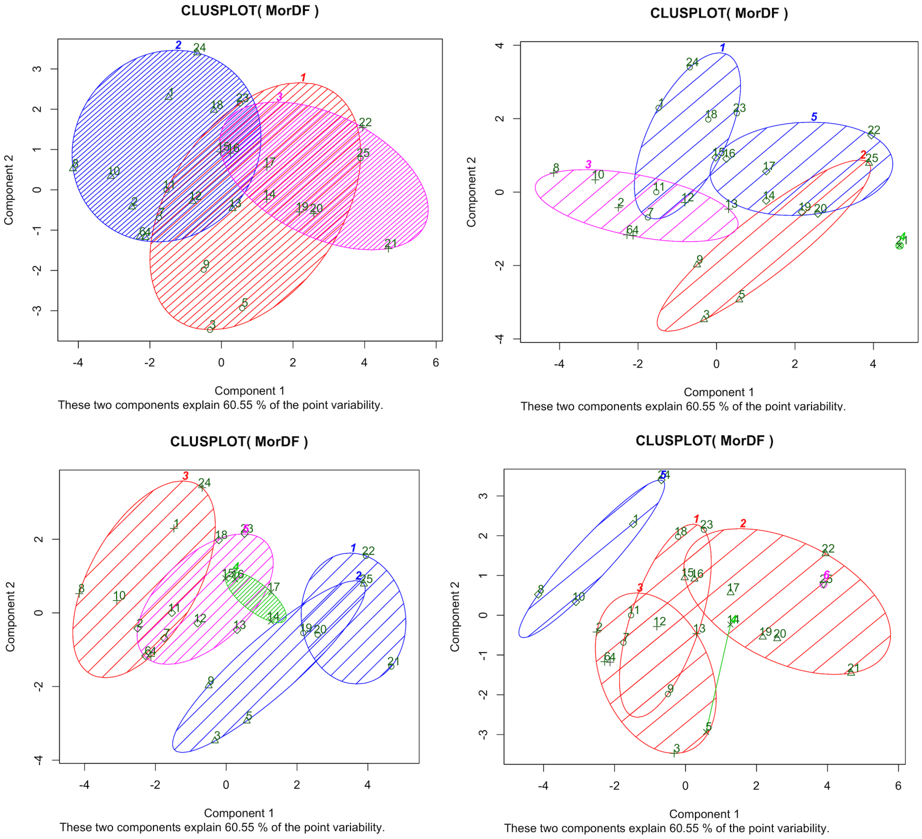
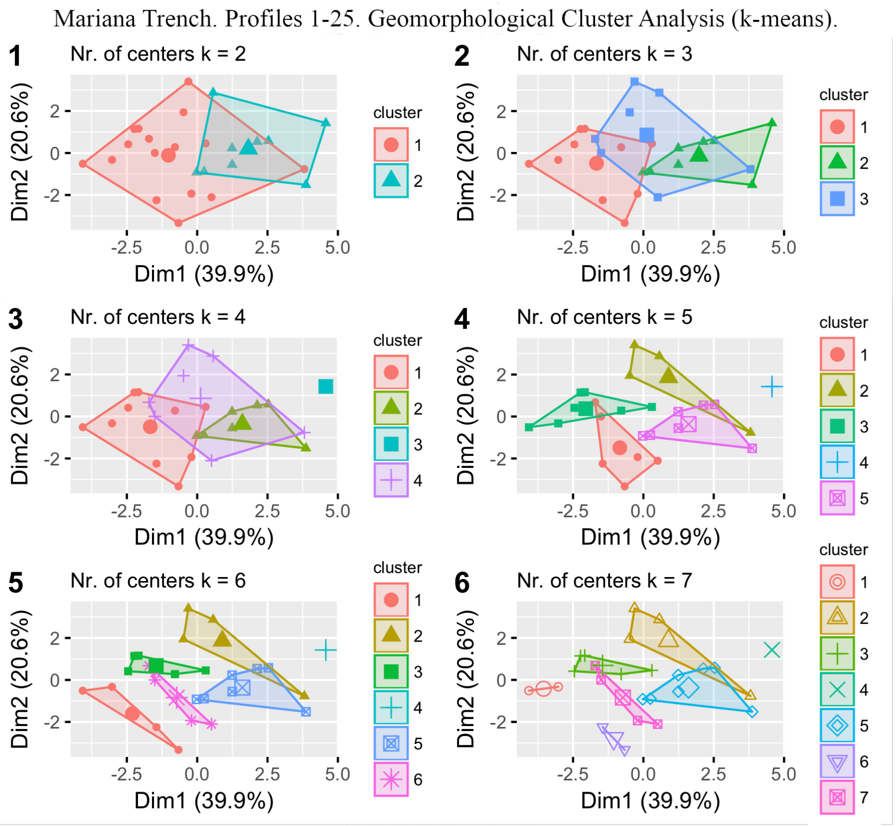


Fig. 6. Testing cluster groups: 3 to 7 of the data set. Computed and visualized in R.

The results of the *k*-means clustering are shown on the Fig. 6 and Fig. 7, the correlation matrix is presented on Fig. 9. The pairwise standard scatter plot of *k*-means cluster correlation distributed by four tectonic plates was done using R code shown on Listing 1.6 with the results on Fig. 10.

Advantages of the k-means Clustering in Data Analysis in Marine Geology The advantages of the k-means clustering among other types of the clustering techniques, e.g., [8,44,5,30], applied for the Mariana Trench consists in the algorithm nature: rather than defining groups before analyzing data, clustering enabled to find and model groups in the profiles that formed organically. Being an unsupervised machine learning algorithm, a *k*-means is an effective and objective algorithm for quantitative and qualitative data analysis, free from the human possible biasses. As performed in the analysis of data distribution, the data were tested on their normality [32,36]. Since the structure of the tectonics and trench geomorphic properties are rather complex, clustering facilitated data partition and grouping. Clusters (Fig. 7) graphically illustrate the results of the *k*-means clustering performed by the R packages {factoextra} and {cluster}.



Statistics Processing and Graphs:
R Programming. Data Source: QGIS

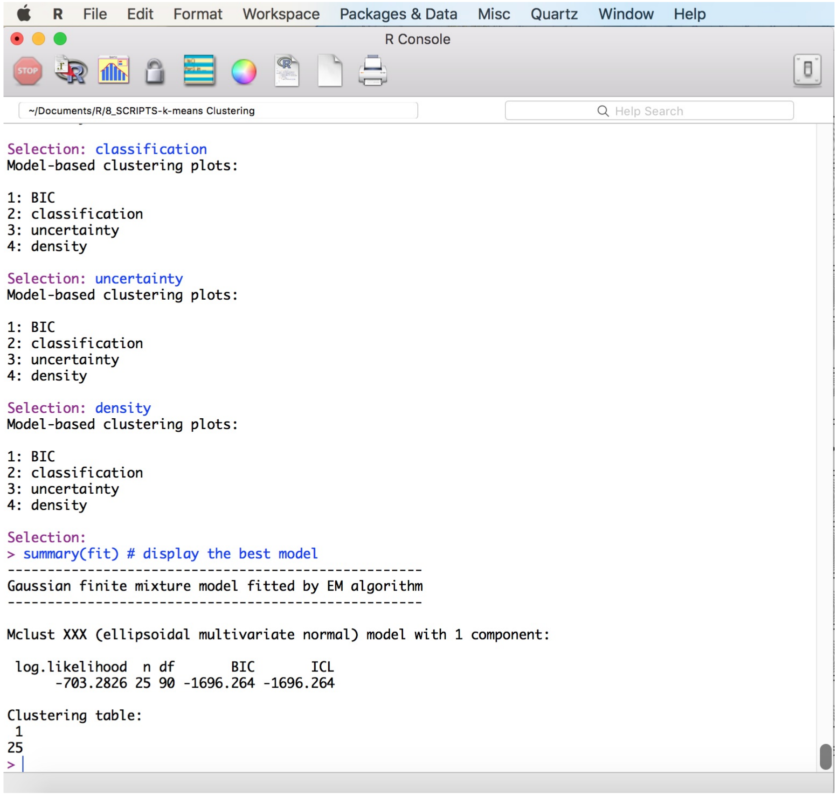
Fig. 7. Results of the k-means clustering of the Mariana Trench with different *k*-values. Computed and visualized in R.

Similarity increases in each class with centroids from k=2 to k=7. Because the nature of the cluster analysis algorithm consists in the iterative process of the discovery of the optimal classes, that is interactive multi-objective optimization, a set of the trial processes was performed before reaching a final decision. The optimal number of the highlighted clusters was finally specifies as five. The workflow opted for five clusters as this presents the optimal compromise between the complexity of the actual data set and their visualization. Samples were then divided into five distinct classes grouped according to their geomorphological similarity of the bathymetric shapes by the cross-section profiles.

```r
library(mclust)
fit <- Mclust(MorDF)
plot(fit)
# plot results
summary(fit)
```

Listing 1.7. R code for the model fitting using {mclust} R program shown on Fig. 8

The sixth class was the smallest class that was further sorted into two subclasses, indicated by the green line (Fig. 6). Therefore, increasing further cluster groups was not necessary, as the optimal number was reached.



```
Selection: classification
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

Selection: uncertainty
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

Selection: density
Model-based clustering plots:

1: BIC
2: classification
3: uncertainty
4: density

Selection:
> summary(fit) # display the best model
----------------------------------------------------
Gaussian finite mixture model fitted by EM algorithm
----------------------------------------------------

Mclust XXX (ellipsoidal multivariate normal) model with 1 component:

 log.likelihood  n df       BIC       ICL
      -703.2826 25 90 -1696.264 -1696.264

Clustering table:
 1
25
> 
```

Fig. 8. Printscreen of R process of identifying related components in Gaussian finite mixture model for clustering

R implements several approaches and algorithms for clustering data frames in {factoextra} and and {cluster} libraries, including the k-means algorithm that was tested in the scope of this research. The aim of cluster analysis is to divide a data frame into significantly distinct groups, or clusters. In this research, the observations, after several test trials were divided into five clusters as optimal number 5 (Fig. 6, lower left sub-plot). These clusters correspond with the observed groupings of the consecutive cross-section profiles containing observation samples (points with geographic XY coordinates and geologic attributes).

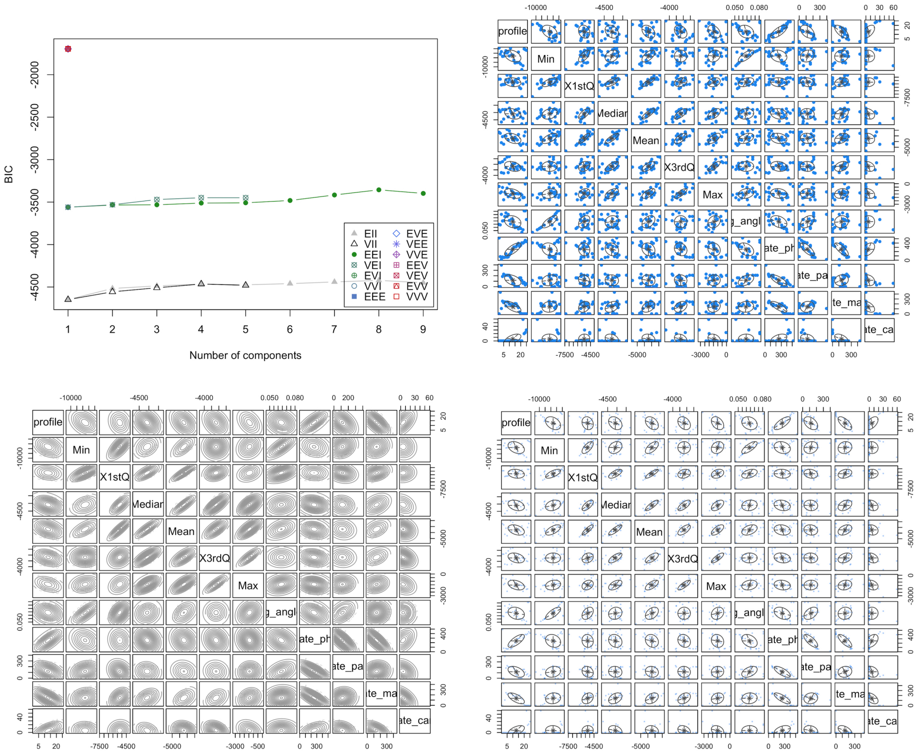


Fig. 9. Correlation matrix, computed and visualized in R. Assessment of the fit versus model complexity using BIC.

The clusters indicate significant geomorphic variations in the geopatial data pool of the Mariana Trench crossing four tectonic plates: Pacific, Philippine Sea, Caroline and Mariana [28]. Assessment of the fit versus model complexity of the clustering was done using using BIC, Fig. 8. Both BIC and Akaike Information Criterion (AIC) are statistical criteria for model selection among a finite set of models based on the likelihood function. Both BIC and AIC attempt to resolve the problem of the model overfitting which happens when adding model parameters aimed at increasing the likelihood of the mode. Comparing BIC and AIC, both deal with the overfitting of the model, introducing an allowed level for the number of parameters in the model. However, the BIC is more effective than AIC, since it reduces the complexity of the model where it refers to the number of parameters. Therefore, the BIC was used for the criteria of model selection using code in Listing 1.7 with a print screen of the process on Fig. 8 and result output on Fig. 9.

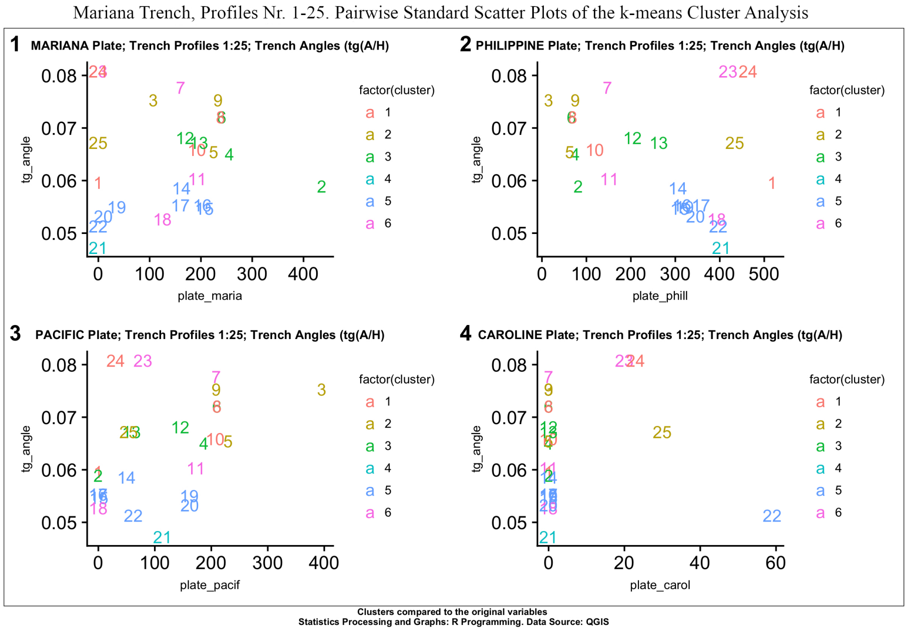


Fig. 10. Cluster groups with number of observation points according to their distributions by four tectonic plates.

3.3 Principal Component Analysis

A statistical analysis based on the orthogonal transformation aimed to convert a set of the depth observations of correlated variables, has been performed using PCA. The PCA (Fig. 10) enabled to visualize eigenvectors showing major direction and vector length for the principal components affecting the categorical values: bathymetry. The direction of the eigenvectors shows the depth values of the Mariana trench along the 25 profiles influenced by the geologic settings and location as well as the similarities among the profiles. the PCA analysis has been performed using R code provided in Listing 1.8. The PCA enables to understand how the variables of the bathymetric data set are varying from the mean depths with respect to each other and if there are relationship between them.

```
# Principal Component Analysis (PCA). libraries: ’factoextra’, ’FactoMiner’ ’zip’, ’openxlsx’, ’carData’, ’pbkrtest’, ’rio’, ’car’, ’flashClust’, ’leaps’, ’scatterplot3d’, ’FactoMineR’, ’ca’, ’igraph’
# Part 1 Creating dataframe.
MDepths <- read.csv("Depths.csv", header=TRUE, sep = ",") # Reading Table.
df<- read.csv("Depths.csv", header=TRUE, sep = ",")
MDF<- na.omit(MDepths) # step-2. cleaning dataframe from NA values
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))})
sum(row.has.na) # count NAs: [1] 0
head(MDF) # ready-to-use dataframe.
# Part 2. Create plot of Principal Component Analysis (PCA)
PCA_Mariana <- autoplot(prcomp(MDF), loadings = TRUE, loadings.colour = ’blue’, loadings.label = TRUE, loadings.label.size = 3) +
  geom_point(color = "blue") +
  scale_color_brewer(palette="Dark2") +
  labs(title="Mariana Trench, Profiles Nr.1-25.",
  subtitle = "PCA (Principal Component Analysis)",
  caption = "Statistics Processing and Graphs: \nR Programming. Data Source: QGIS") +
  theme() # Legend design
PCA_Mariana
```

Listing 1.8. R code for the PCA

Variables in the groups of the profiles are highly correlated: 1) group 1 (Nr. 4, 15,16); 2) groups 2 (Nr. 10, 20, 22); 3) group 3 (profiles 19 and 9); 4) group 4 (Nr. 25, 5, 12); 5) group 5 (Nr. 3, 1, 24); 6) group 6 (Nr. 6, 17, 4). Other profiles have more individual shape with a clear distinction of the profile Nr. 21 where the deepest samples are recorded.

4 Results

Findings in correlation analysis and the results of the k-means algorithm clustering and data grouping show groups across all 25 profiles, with the number of groups represented by the variable. Several possible clusters were tested from two to seven. It was found out that the optimal number is five: in this case, the cluster circles contain the optimal number of the observations and the overlapping was reasonably minimal (Fig. 6). The correlation matrix is presented on Fig. 9 showing crossing correlations in the combination of the environmental factors. Comparison of the *bi*-factor in-between the factors revealed pairwise correlation (Fig. 10).

Pairwise comparative analysis (Fig. 10) enabled to observe a marked influence on the environmental variables as *bi*-factors. Thus, in response to the decreasing sediment thickness the slope angle goes in parallel; location of the volcanic igneous areas cause a cyclic repetition of the curve for the slope angles, as well as those of igneous volcanic areas have certain correlation between the slope angle and aspect degree. Therefore, according to the findings, four environmental variables are affecting the geomorphological structure of the trench. These include slope angle, sediment thickness, aspect degree and location of the volcanic igneous areas.

The summary of the results is as follows:

- correlation matrix showing crossing correlations in the combination of factors;
- comparison of the *bi*-factors in-between the factors revealed pairwise correlation;
- pairwise comparative analysis enabled to observe an influence on the variables as *bi*-factors: in response to the decreasing sediment thickness, slope angles go in parallel;
- the location of the volcanic igneous areas correlate with the slope angles, while volcanic zones correlate with the slope angle and aspect degree.

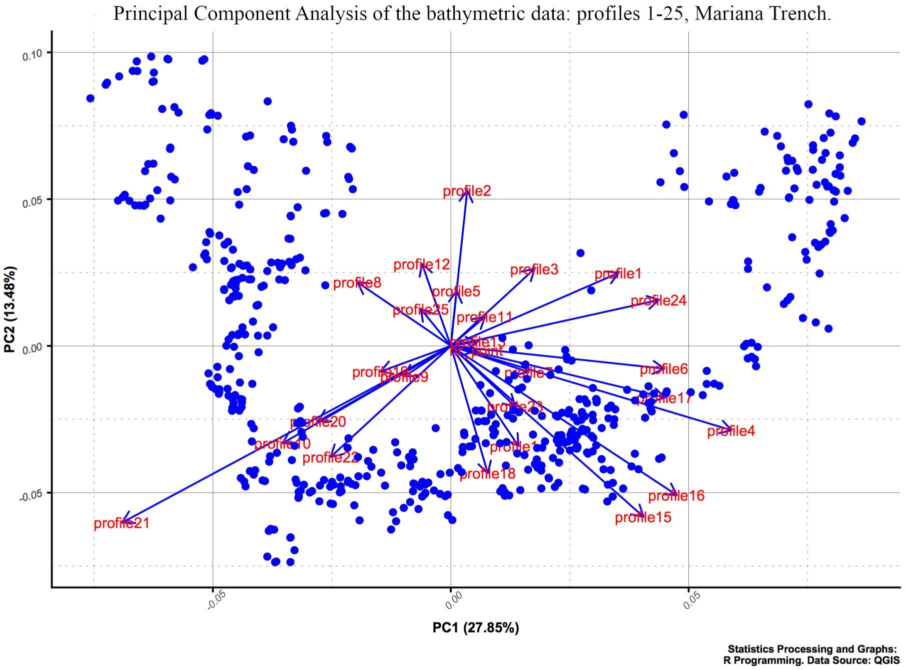


Fig. 11. Principal Component Analysis of the bathymetric data by profiles.

5 Discussion

Many studies described the environmental settings at extreme depths of the Pacific Ocean, to mention some of them, e.g. [14,15,17,37,45]. As pointed by [16], modern studies in the marine biology are currently limited to the bathyal

(2003,000m) and abyssal depths (3,0006,000m). The shape of the trench geomorphology affects all aspects of the marine ecosystems. The specific features of the benthic biology consists in its significance: although it covers 12% of the global benthic area, the hadal zone constitutes the deepest 45% area of the vertical depth gradient [19]. Hence, the deep-sea trenches represent a cluster of the complex geomorphology that include features of the continental slopes, abyssal plains and unique properties of the geomorphic shapes affecting the environmental aspects of the hadal ecosystems. Life in the hadal trenches is strongly restricted by a variety of factors that delimit vertical constrains within distinct bathymetric strata and define the distribution of the marina organisms.

The effects of the hydrostatic pressure, temperature, salinity, oxygen and food supply strongly affect and determine the location of the species. Thus, the pressure within the trench increases by 1 MPa per 100 m, reaching 100 MPa in the deepest places of the trenches [50]. This well illustrates the complexity of the marine environment notable for the deep-sea life conditions. At the same time, so far the understanding of how the life in functioning in such remotely located areas is not sufficient.

Precise analysis of the data sets on the deep-sea marine ecosystems that include a variety of factors with interrelated attributes is only possible by means of the machine learning. Besides statistical methods of data analysis, such as SPSS Statistics [25], Gretl statistical software [27] Python libraries [4,26,29], such functionality is fully possible by means of R programming. A contemporary perspective of the statistical methods for the analysis of trench structure and formation is demonstrated in this research. The application of the k-means clustering method provided by the functionality of R programming offers optimal prospects for better understanding of the bathymetry of the deep ocean trenches. The k-means clustering enables to test groups of the observation data for further geostatistical processing using embedded techniques in R libraries {cluster} and {factoextra}.

Clustering technique, as demonstrated in this research, is associated with the spatial distribution analysis of the sample points. The k-means clustering method considers spatial correlation between the samples belonging to the same group and statistical relationships between the observation points within the data set. A rigorous and quantitative clustering analysis performed by R is an effective tool for the geological investigation of such complex structures as deep-sea trenches. Hence, current work contributes towards the development of the technical methods of the statistical analysis by means of R programming applied to the geological data sets.

This paper demonstrated an example of the cross-disciplinary quantitative approach for the geological analysis with an R scripting approach. Modelling data by clustering analysis using R does not only make geological modelling simpler and less error-prone. It may also facilitate more complex simulations involving, for instance, multi-dimensional modelling that runs with varying geological parameters. Before any complex modelling, a data analysis such as data distribution and grouping by clustering is important, as demonstrated in this paper.

Current research provides R codes used for plotting, what makes possible to apply these methods for testing in similar research where data analysis by k-means clustering is required.

6 Acknowledgements

The study is supported by China Scholarship Council (CSC), State Oceanic Administration (SOA), Marine Scholarship of China, Grant Nr. 2016SOA002, Beijing, People's Republic of China (P.R.C.)

7 Acronyms

List of notations and acronyms [1]

AIC	Akaike Information Criterion	17
BIC	Baysiean Information Criterion	12
DF	Data Frame	5
GMT	Generic Mapping Tools	2
QGIS	Quantum GIS	2
IQR	InterQuartile Range	7
IT	Information Technologies	2
k	k centers in clustering algorithm	5
MorDF	Morphology data frame	13
MDepths	Marina trench depth values data frame	5
NA	Non available (numbers)	5
Nr	Number	10
PCA	Principal Component Analysis	5
SRTM	Shuttle Radar Topography Mission	4
UTM	Universal Transverse Mercator	4

[1] The page is given where the glossary is first entered and defined

References

1. Agarwal, P.K., Procopiuc, C.M.: Exact and approximation algorithms for clustering. In: Proceedings of the 9th Annual ACM-SIAM Symposium on Discrete Algorithms, San Francisco, CA. pp. 658–667 (1998)
2. Arthur, D., Vassilvitskii, S.: k-means++: the advantages of careful seeding. In: Proceedings of the 18th annual ACM-SIAM symposium on Discrete algorithms. Society for Industrial and Applied Mathematics Philadelphia. pp. 1027–1035 (2007)
3. Ciaccio, A.D., Coli, M., Angulo Ibanez, J.M.: Studies in Theoretical and Applied Statistics Selected Papers of the Statistical Societies, chap. Advanced Statistical Methods for the Analysis of Large Data Sets. Springer (2012). https://doi.org/10.1007/978-3-642-21037-2
4. Cielen, D., Meysman, A.D.B., M., A.: Introducing Data Science. Big Data, Machine Learning and More, Using Python Tools. Manning, Shelter Island, U.S. (2016)
5. Deng, X.H., Chen, Y.J., Pirajno, F., Li, N., Yao, J.M., Sun, Y.L.: The geology and geochronology of the Waifangshan Mo-quartz vein cluster in eastern Qinling, China. Ore Geology Reviews **81**, 548–564 (3 2017). https://doi.org/10.1016/j.oregeorev.2015.10.009
6. Dumont, M., Reninger, P.A., Pryet, A., Martelet, G., Aunay, B., Join, J.L.: Agglomerative hierarchical clustering of airborne electromagnetic data for multi-scale geological studies. Journal of Applied Geophysics **157**, 1–9 (10 2018). https://doi.org/10.1016/j.jappgeo.2018.06.020
7. Efendiyev, G.M., Mammadov, P.Z., Piriverdiyev, I.A., Mammadov, V.N.: Clustering of Geological Objects Using FCM-algorithm and Evaluation of the Rate of Lost Circulation. Procedia Computer Science **102**, 159–162 (2016). https://doi.org/10.1016/j.procs.2016.09.383
8. Efendiyev, G.M., Rza-zadeh, S.A., Kadimov, A.K., Kouliyev, I.R.: Forecast of drilling mud loss by statistical technique and on the basis of a fuzzy cluster analysis. In: Proceedings of the 7th International Conference on Soft Computing, Computing with Words and Perceptions in System Analysis, Decision and Control. pp. 319–322 (2013)
9. Fabbrocino, S., Rainieri, C., Paduano, P., Ricciardi, A.: Cluster analysis for groundwater classification in multi-aquifer systems based on a novel correlation index. Journal of Geochemical Exploration **204**, 90–111 (2019). https://doi.org/10.1016/j.gexplo.2019.05.006
10. Faber, V.: Clustering and the continuous k-means algorithm. Los Alamos Science **22**, 138–144 (1994)
11. Fan, Z., Xu, X.: Application and visualization of typical clustering algorithms in seismic data analysis. Procedia Computer Science **151**, 171–178 (2019). https://doi.org/10.1016/j.procs.2019.04.026
12. Farr, T.G., Rosen, P.A., Caro, E., Crippen, R., Duren, R., Hensley, S., Kobrick, M., Paller, M., Rodriguez, E., Roth, L., Seal, D., Shaffer, S., Shimada, J., Umland, J., Werner, M., Oskin, M., Burbank, D., Alsdorf, D.: The Shuttle Radar Topography Mission. AGU Review of Geophysics **45**, 1–33 (2007). https://doi.org/10.1029/2005RG000183
13. Foroud, T., Seifi, A., AminShahidy, B.: An efficient optimization process for hydrocarbon production in presence of geological uncertainty using a clustering method: A case study on Brugge field. Journal of Natural Gas Science and Engineering **32**, 476–490 (5 2016). https://doi.org/10.1016/j.jngse.2016.04.059

14. Hartwell, A.M., Voight, J.R., Wheat, C.G.: Clusters of deep-sea egg-brooding octopods associated with warm fluid discharge: An ill-fated fragment of a larger, discrete population? Deep-Sea Research Part I: Oceanographic Research Papers **135**, 1–8 (2018). https://doi.org/10.1016/j.dsr.2018.03.011
15. Hessler, R.R., Ingram, C.L., Yayanos, A.A., Burnett, B.: Scavenging amphipods from the floor of the Philippine Trench. Deep-Sea Research Part I: Oceanographic Research Papers **25**, 1029–1047 (1978). https://doi.org/10.1016/0146-6291(78)90585-4
16. Ichino, M.C., Clark, M.R., Drazen, J.C., Jamieson, A., Jones, D.O.B., Martin, A.P., Rowden, A.A., Shank, T.M., Yancey, P.H., Ruhl, H.A.: The distribution of benthic biomass in hadal trenches: A modelling approach to investigate the effect of vertical and lateral organic matter transport to the seafloor. Deep-Sea Research Part I: Oceanographic Research Papers **100**, 21–33 (2015). https://doi.org/10.1016/j.dsr.2015.01.010
17. Itoh, M., Kawamura, K., Kitahashi, T., kiKojima, S., Katagiri, H., Shimanaga, M.: Bathymetric patterns of meiofaunal abundance and biomass associated with the Kuril and Ryukyu trenches, western North Pacific Ocean. Deep-Sea Research Part I: Oceanographic Research Papers **58**, 86–97 (2011). https://doi.org/10.1016/j.dsr.2010.12.004
18. Jamieson, A.J., Fujii, T.: Trench Connection. Biology Letters **7**, 641–643 (2011). https://doi.org/10.1098/rsbl.2011.0231
19. Jamieson, A.J., Fujii, T., Mayor, D.J., Solan, M., Priede, I.G.: Hadal trenches: the ecology of the deepest places on Earth. Trends in Ecology and Evolution **25**(3), 190–197 (2009). https://doi.org/10.1016/j.tree.2009.09.009
20. Lee, K., Jung, S., Choe, J.: Ensemble smoother with clustered covariance for 3D channelized reservoirs with geological uncertainty. Journal of Petroleum Science and Engineering **145**, 423–435 (09 2016). https://doi.org/10.1016/j.petrol.2016.05.029
21. Lemenkova, P.: Factor Analysis by R Programming to Assess Variability Among Environmental Determinants of the Mariana Trench. Turkish Journal of Maritime and Marine Sciences **4**, 146–155 (12 2018). https://doi.org/10.6084/m9.figshare.7358207
22. Lemenkova, P.: Hierarchical Cluster Analysis by R language for Pattern Recognition in the Bathymetric Data Frame: a Case Study of the Mariana Trench, Pacific Ocean. In: Krasnyansky, M.N. (ed.) Proceedings of the 5th International Conference 'Virtual Simulation, Prototyping and Industrial Design'. vol. 2, pp. 147–152. TSTU Press (2018). https://doi.org/10.6084/m9.figshare.7531550
23. Lemenkova, P.: R scripting libraries for comparative analysis of the correlation methods to identify factors affecting Mariana Trench formation. Journal of Marine Technology and Environment **2**, 35–42 (11 2018). https://doi.org/10.6084/m9.figshare.7434167
24. Lemenkova, P.: An Empirical Study of R Applications for Data Analysis in Marine Geology. Marine Science and Technology Bulletin **8**, 1–9 (03 2019). https://doi.org/10.33714/masteb.486678
25. Lemenkova, P.: Numerical Data Modelling and Classification in Marine Geology by the SPSS Statistics. International Journal of Engineering Technologies **5**, 90–99 (06 2019). https://doi.org/10.6084/m9.figshare.8796941
26. Lemenkova, P.: Processing oceanographic data by Python libraries NumPy, SciPy and Pandas. Aquatic Research **2**, 73–91 (04 2019). https://doi.org/10.3153/AR19009

27. Lemenkova, P.: Regression Models by Gretl and R Statistical Packages for Data Analysis in Marine Geology. International Journal of Environmental Trends **3**, 39–59 (06 2019). https://doi.org/10.6084/m9.figshare.8313362.v1
28. Lemenkova, P.: Scatterplot Matrices of the Geomorphic Structure of the Mariana Trench at Four Tectonic Plates (Pacific, Philippine, Mariana and Caroline): a Geostatistical Analysis by R. In: Degtyarev, K.E. (ed.) Problems of Tectonics of Continents and Oceans. vol. 1, pp. 347–352. Institute of Geology, Russian Academy of Science, GEOS (2019). https://doi.org/10.6084/m9.figshare.7699787.v1
29. Lemenkova, P.: Testing Linear Regressions by StatsModel Library of Python for Oceanological Data Interpretation. Aquatic Sciences and Engineering **34**, 51–60 (06 2019). https://doi.org/10.26650/ASE2019547010
30. Martin, R., Boisvert, J.: Towards justifying unsupervised stationary decisions for geostatistical modeling: Ensemble spatial and multivariate clustering with geomodeling specific clustering metrics. Computers and Geosciences **120**, 82–96 (11 2018). https://doi.org/10.1016/j.cageo.2018.08.005
31. Michel, V., Gramfort, A., Varoquaux, G., Eger, E., Keribin, C., Thirion, B.: A supervised clustering approach for fMRI-based inference of brain states. Pattern Recognition **45**, 2041–2049 (6 2012). https://doi.org/10.1016/j.patcog.2011.04.006
32. Myers, J.L., Well, A.D.: Research Design and Statistical Analysis. Lawrence Erlbaum, 2 edn. (2003)
33. R Development Core Team: R: a language and environment for statistical computing. R Foundation, Vienna, Austria (2014), http://www.R-project.org, last accessed 30 Jul 2019
34. Roberts, N.M., Tikoff, B., Davis, J.R., Stetson-Lee, T.: The utility of statistical analysis in structural geology. Journal of Structural Geology **125**, 64–73 (8 2019). https://doi.org/10.1016/j.jsg.2018.05.030, reference: SG 3671; PII: S0191-8141(17)30339-5
35. Romankevich, E.A., Vetrov, A.A., Peresypkin, V.I.: Organic matter of the World Ocean. Russian Geology and Geophysics **50**, 299–307 (2008). https://doi.org/10.1016/j.rgg.2009.03.013
36. Marques de Sá, J.P.: Applied Statistics Using SPSS, Statistics, Matlab and R. Springer, Porto, Portugal, 2 edn. (2007)
37. Stewart, H.A., Jamieson, A.J.: Habitat heterogeneity of hadal trenches: Considerations and implications for future studies. Progress in Oceanography **161**, 47–65 (2018). https://doi.org/10.1016/j.pocean.2018.01.007
38. Swan, A.R.H., Sandilands, M.: Introduction to Geological Data Analysis. Blackwell Science, Cambridge, Massachusetts, USA (1995)
39. Szabó, N.P., Nehéz, K., Hornyák, O., Piller, I., Deák, C., Hanzelik, P.P., Kutasi, C., Ott, K.: Cluster analysis of core measurements using heterogeneous data sources: An application to complex Miocene reservoirs. Journal of Petroleum Science and Engineering **178**, 575–585 (2019). https://doi.org/10.1016/j.petrol.2019.03.067
40. Tian, J., Fan, L., Liu, H., Liu, J., Li, Y., Qin, Q., Gong, Z., Chen, H., Sun, Z., Zou, L., Wang, X., Xu, H., Bartlett, D., Wang, M., Zhang, Y.Z., Zhang, X.H., Zhang, C.: A nearly uniform distributional pattern of heterotrophic bacteria in the Mariana Trench interior. Deep-Sea Research Part I: Oceanographic Research Papers **142**, 116–126 (2018). https://doi.org/10.1016/j.dsr.2018.10.002
41. Van Haren, H., Berndt, C., Klaucke, I.: Ocean mixing in deep-sea trenches: New insights from the Challenger Deep, Mariana Trench. Deep-Sea Research Part I: Oceanographic Research Papers **129**, 1–9 (11 2017). https://doi.org/10.1016/j.dsr.2017.09.003

42. Vermeesch, P., Resentini, A., Garzanti, E.: An R package for statistical provenance analysis. Sedimentary Geology **336**, 14–25 (2016). https://doi.org/10.1016/j.sedgeo.2016.01.009
43. Wang, J., Zuo, R., Caers, J.: Discovering geochemical patterns by factor-based cluster analysis. Journal of Geochemical Exploration **181**, 106–115 (2017). https://doi.org/10.1016/j.gexplo.2017.07.006
44. Wang, R., Wang, Z., Osumanu, A., Zhang, G., Li, B., Lu, Y.: Grid density overlapping hierarchical algorithm for clustering of carbonate reservoir rock types: A case from Mishrif Formation of West Qurna-1 oilfield, Iraq. Journal of Petroleum Science and Engineering **182**, 106–209 (2019). https://doi.org/10.1016/j.petrol.2019.106209
45. Webb, T.J., Berghe, E.V., O'Dor, R.: Biodiversitys Big Wet Secret: The Global Distribution of Marine Biological Records Reveals Chronic Under-Exploration of the Deep Pelagic Ocean. PlosOne **5**, 1–6 (8 2010). https://doi.org/10.1371/journal.pone.0010223
46. Wessel, P., Smith, W.H.F.: A Global Self-consistent, Hierarchical, High-resolution Shoreline Database. Journal of Geophysical Research Atmospheres **101**, 8741–8743 (1996). https://doi.org/10.1029/96JB00104
47. Wessel, P., Smith, W.H.F.: New version, of the Generic Mapping Tools released. EOS Transactions of the American Geophysical Union **79**(47), 329 (1998). https://doi.org/10.1029/98EO00426
48. Wessel, P., Smith, W.H.F.: The Generic Mapping Tools. Version 4.5.18 Technical Reference and Cookbook. GMT, U.S.A (2018)
49. Xu, Y., Ge, H., Fang, J.: Biogeochemistry of hadal trenches: Recent developments and future perspectives. Deep-Sea Research Part II: Topical Studies in Oceanography **155**, 19–26 (2018). https://doi.org/10.1016/j.dsr2.2018.10.006
50. Yancey, P.H., Gerringer, M.E., Drazen, J.C., Rowden, A.A., Jamieson, A.: Marine fish may be biochemically constrained from inhabiting the deepest ocean depths. PNAS (Proceedings of the National Academy of Sciences of the United States of America) **111**, 4461–4465 (2014). https://doi.org/10.1073/pnas.1322003111