Hierarchical Cluster Analysis by R language for Pattern Recognition in the Bathymetric Data Frame: a Case Study of the Mariana Trench, Pacific Ocean

Polina Lemenkova

УДК 551.46

Леменкова Полина Алексеевна[1]
[1] Океанологический университет Китая, Колледж морских геонаук, КНР, г.Циндао
(Тел. (+86)1768-554-1605 e-mail: pauline.lemenkova@gmail.com)

ИЕРАРХИЧЕСКИЙ КЛАСТЕРНЫЙ АНАЛИЗ НА ЯЗЫКЕ R ДЛЯ ВЫДЕЛЕНИЯ ОДНОРОДНЫХ ГРУПП БАТИМЕТРИЧЕСКИХ ДАННЫХ НА ПРИМЕРЕ МАРИАНСКОГО ЖЕЛОБА, ТИХИЙ ОКЕАН

Lemenkova Polina[1]
[1]Ocean University of China, College of Marine Geoscience, P.R.C., Qingdao
(Tel. (+86)1768-554-1605 e-mail: pauline.lemenkova@gmail.com)

HIERARCHICAL CLUSTER ANALYSIS BY R LANGUAGE FOR PATTERN RECOGNITION IN THE BATHYMETRIC DATA FRAME: A CASE STUDY OF THE MARIANA TRENCH, PACIFIC OCEAN

Аннотация: Целью настоящего исследования является анализ батиметрических свойств Марианского желоба, самой глубокой точки Земли, расположенной в западной части Тихого океана. Марианский желоб имеет уникальную структуру и особенности, сформированные в сложном процессе его развития под влиянием ряда геолого-тектонических факторов. Батиметрические свойства поперечных профилей желоба сильно различаются в разных его частях: угол и крутизна склона, глубина, покатость, выположенность. Технически, работа опиралась на использование статистических алгоритмов кластерного анализа группировки данных на языке программирования R. Визуализация и получение данных были осуществлены в QuantumGIS. Результаты выявили неравномерность батиметрических профилей согласно факторам, влияющим на геологическую структуру желоба.

Ключевые слова: кластерный анализ, R, программирование, алгоритмы, океанография, Марианский желоб, Тихий океан

Abstract. The geographic focus of the current study Mariana trench, the deepest point of the Earth located in the west Pacific Ocean. Mariana trench has unique structure and features formed in the complex process of the trench development. There is a range of the environmental factors affecting trench structure and functioning: bathymetry, geography, geology and tectonics. Current research aimed to study interconnections among these determinants. Technically, the research was performed by R programming language, statistical analysis, and QuantumGIS. Methodology includes a range of the statistical methods for data processing, the most important of which is cluster analysis. The results revealed unevenness of the factors affecting trench bathymetric structure, caused by the environmental conditions.

Keywords: Cluster analysis, R, programming, algorithms, oceanography, Mariana Trench, Pacific Ocean

1. Introduction

Mariana trench has a complicated geomorphic structure with various types of the continental tectonic plate boundaries producing a unique type of its seafloor bathymetry (Fig.1). Submarine terrain of the Mariana trench has a complex structure of various objects, the scale of which sharply exceeds the dimensions forms of the earth's surface on the margin continents. These include the system of the mid-ocean ridges, crossing their sub-system of transform faults, deep-sea ridges, vast ocean plains complicated by chains and mountain groups, ridges and plateaus [3]. The width of the

continental shelf of the margin plates connecting to the Mariana trench occupies varies from a few tens to thousands of km and the average bottom slope is 17-25° [1].

The main ridge axes of the Mariana trench represent a place in the system where the upwelling magma encounters cold seawater in the location of four tectonic plates crossing Mariana trench. Since functionality of the biological communities of the Marina trench goes beyond the scopes of the current research, it should be only briefly mention as an example of interconnection of the system [2], [4]. Hence, the hydrothermal cooling of magma at the ridge axes causes recycling of the entire volume of the water in the hadal column. In this way it provides nutrients for biological communities in the abyssal of trench, as studies, for example in various research papers [5], [6], [7].

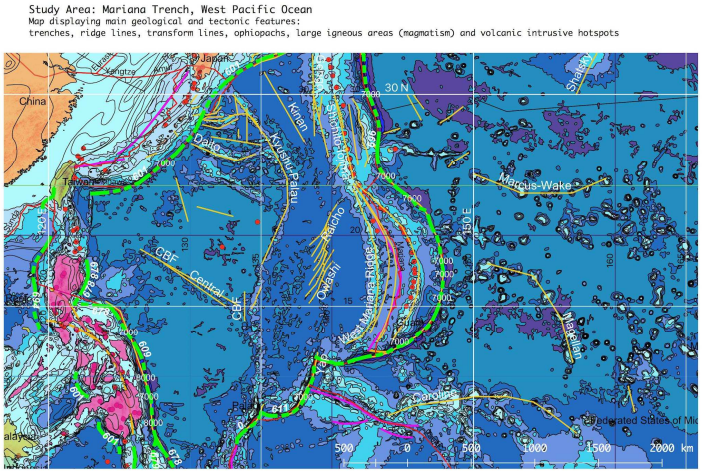

Fig. 1. Mariana Trench: QuantumGIS visualization

2. Research Problem

Systematic ocean exploration in the 21th century started combining remote sensing methods, GIS tools for data processing and a variety of machine learning tools for data post-processing and statistical analysis. The study of the ocean floor is possible not only through the GIS, yet also by means of the available R statistical approaches: network analysis, factors analysis, plotting, diagrams, cluster analysis by various methods (k-means, dendrograms, silhouettes), triple correlation by ternary diagrams, to mention a few.

3. Methods and Results

In the current research, the main method of the determining clusters of the environmental variables affecting Mariana trench formation is an algorithm provided by cluster analysis, executed by {dendextend} R library. Initially, the GIS based digitizing of the 25 bathymetric profiles across Mariana trench has been performed in QGIS. The length of each profile was taken 1000 km, and the distance between every pair was 100 km (Fig.2). The coordinated were saved in a table with three columns: elevations, latitude and longitude. After that, a cycle number of the R codes enabled to stepwise perform cluster analysis, data partition, sorting and grouping. Comparative analysis of the multi-dimensional data enabled to analyse inter-relationships between the factors impacting changes in the morphology and dynamics of the system of Mariana trench.

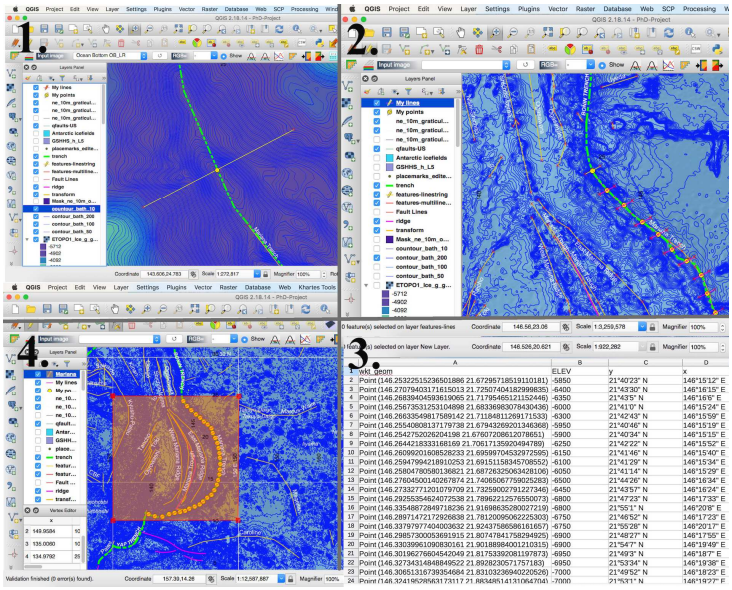

Fig. 2. Four steps of creating 25 bathymetric profiles in QuantumGIS

Using existing methodologies, various approaches of the comparative analysis were tested in the current research. The most efficient was computing hierarchical dendrogram clustering with p-values called by R library {dendextend}. The default hierarchical clustering method used in this library is "hclust". The results of the executed algorithm have been visualized as a plotted dendrogram.

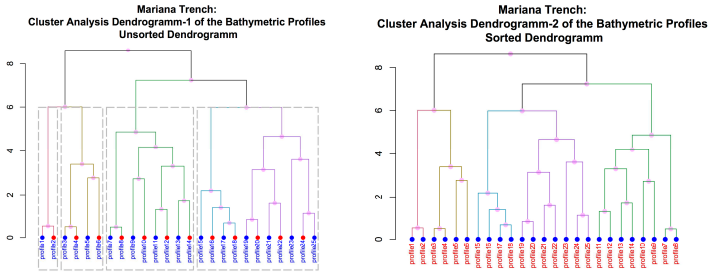

Fig. 3. R based dendrogram tree of the 25 profiles. Left: unsorted, right: sorted

The working flowchart include following steps: creating unsorted and sorted dendrograms (Fig. 3), cross-comparison of the both (Fig. 4), hierarchical clustering using p-Values, and dendrogram by clustering using bootstrap probability method (Fig. 5). The initial step included created unsorted dendrogram. Several adjustments have been made thereafter: changed and coloured labels based on real profiles groups category, coloured branches based on the procedure of cutting dendrogram tree into the clusters. Then the dendrogram was sorted using machine learning algorithm. After the reflection of the unsorted dendrogram on the sorted according to the data distribution by observation values, the model proceeded to the comparison of the two dendrograms (Fig. 4). The compared dendrograms now form cluster kernels. Further processing is completed on the machine by an embedded mathematical approach of R that filter out the less significant profiles and compares them to the kernel ones.

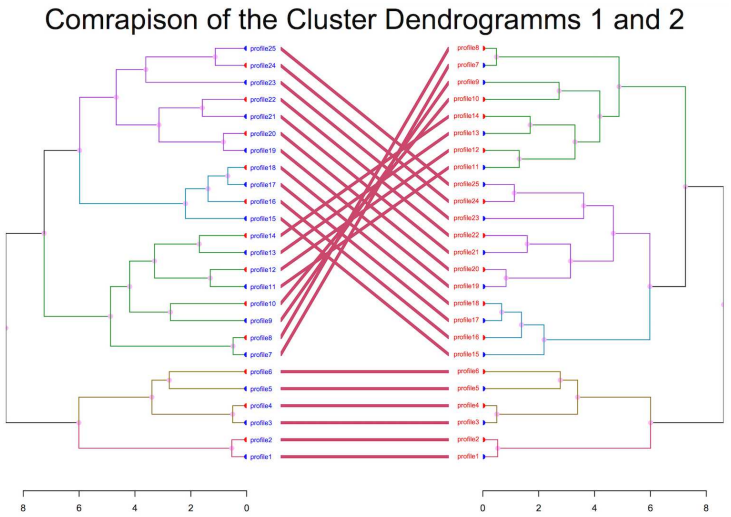

Fig. 4. Cluster analysis in R programming: comparison between cluster groups

The information is then stored to create final step: clustering using bootstrap probability method. Thus, there are four modifications of the initial model consisted of cross-section bathymetric profiles of the Mariana trench. The hierarchical dendrogram method was selected to test the data, due to its flexibility: this type of the machine learning is used specially for unlabelled data, i.e. without defined categories or groups, aimed to understand data possible allocation into groups. As the structure of the tectonics and trench geomorphic properties are rather complex, the dendrogram clustering facilitated data partition though defining the resulting groups, with similar morphology, distance to the volcanic areas, slope angles and sedimental thickness. The complete R programming script for cluster analysis is as follows:

3.1. Creating data.frame

Read-in table, create dataframe as performed using following code :

```
# step-1 MDepths <- read.csv("Mariana.csv", header=TRUE, sep = ",")
# step-2.At the next step the data frame was cleaned up from non-available values:
MDF <- na.omit(MDepths)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))}) # check up if non-available values are deleted
sum(row.has.na) # sum up all NA, should be: [1] 0
head(MDF) # look up cleaned up dataframe.
```

3.2. Hierarchical cluster analysis, dendrogram tree

```
    library(dendextend) – using this package
# step-3. creating 1st dendrogram (here: by 25 clusters as bathymetric profiles)
    dend <- MDF[1:25,] %>%  scale %>% dist %>% # calculate a distance matrix,
    hclust (method = "average") %>%        as.dendrogram %>%
        set("labels", c(("profile"), rep(1:25), sep="")) %>%
        set("labels_col","blue") %>% set("labels_cex", c(.7)) %>%
        set("branches_k_color", k=5) %>% set("branches_lwd", 1) %>%
        set("nodes_pch", 19) %>%  set("nodes_cex", 1) %>%
        set("nodes_col", "plum1") %>%
        set("leaves_pch", 19) %>% set("leaves_col", c("blue", "red"))
    dend %>% plot(main = "Mariana Trench: \nCluster Analysis Dendrogramm-1 of the Bathymetric Profiles \nUnsorted Dendrogramm")
```

```
# step-4. Create the 2nd dendrogram from the 1st one sorted by size of clusters
    dend2 <- sort(dend)
    dend2 %>%  set("branches_k_color", k=3) %>% set("branches_lwd", 1) %>%
      set("labels_col","blue")
    %>% set("labels_cex", c(.7)) %>%      set("branches_k_color",          k=5)          %>%
set("branches_lwd", 1) %>%
      set("nodes_pch", 19) %>%  set("nodes_cex", 1) %>%
      set("nodes_col", "plum1") %>%
      set("leaves_pch", 19) %>% set("leaves_col", c("blue", "red"))
      dend2 %>% plot(main = "Mariana Trench: \nCluster Analysis Dendrogramm-2 of the
Bathymetric Profiles \nSorted Dendrogramm")
```

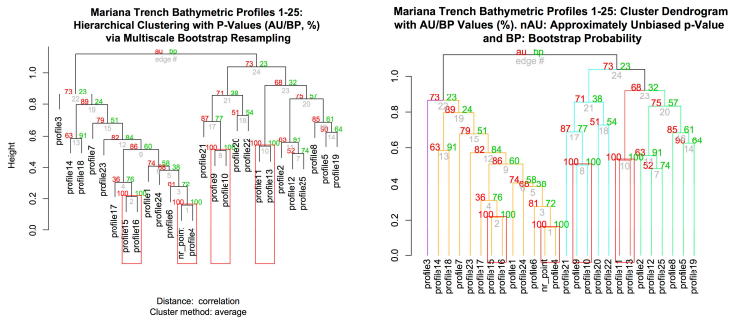

Fig. 5. Hierarchical clustering with p-values using multiscale bootstrap probability

3.3. Comparing dendrograms

```
# step-5. (sorted and unsorted ones) tanglegram(dend, dend2)
    tanglegram(dend, dend2) %>% plot(main = "Mariana Trench: \nComrapison of the Cluster
Dendrogramms 1 and 2")
# step-6. Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling data(MDF)
set.seed(518)
result <- pvclust(MDF, method.dist="cor", method.hclust="average", nboot=10)
# Default plot of the result
plot(result, main = "Mariana Trench Bathymetric Profiles 1-25: \nHierarchical Clustering with P-
Values (AU/BP, %) \nvia Multiscale Bootstrap Resampling")
pvrect(result)
    # step-7. pvclust and dendextend – results as a sorted dendrogram result %>% as.dendrogram
%>%  set("branches_k_color", k = 5, value = c("purple", "orange", "cyan1", "firebrick1",
"springgreen")) %>% plot(main = "Mariana Trench Bathymetric Profiles 1-25: Cluster Dendrogram
nwith AU/BP Values (%). nAU: Approximately Unbiased p-Value \n and BP: Bootstrap
Probability") result %>% text result %>% pvrect
```

4. Conclusion

To summarize methodological application of this research, following conclusion can be drawn. The advantages consist in usability of many R libraries and algorithms at one time to understand the best option for the data set. When the proper method is found for each research step, the results are visualized as plot. In this case, the main method was dendrogram by the {dendextend} library aimed at data partition according to the environmental properties. As the programs received a table

in (.csv) format and called script from the command line, it follows a mathematical algorithm to process the data. The R algorithm then calculated the values for the environmental values (geology, bathymetry, geomorphology, sedimentation and tectonics), and divided the dataset into clusters, according to the similarity of their values. Once the data are analysed by the machine, the visualization of the bathymetric groups and data partition were successfully performed.

References

1. **Пущаровский, Ю.М., Меланхолина Е.Н.** Тектоническое развитие Земли. Тихий океан и его обрамление. – М., – 1992.
2. **Gallo, N.D., Cameron, J., Hardy, K., Fryer, P., Bartlett, D.H., Levin, L.A.** Submersible and lander-observed community patterns in the Mariana and New Britain trenches: Influence of productivity and depth on epibenthic and scavenging communities. Deep-Sea Research I, – 2015 – V. 99. – P.119–133.
3. **Jarrard, R.** (1986). Relations among subduction parameters. Rev. Geoph. – 1986. – V. 24. – P. 217–284.
4. **Lacey, N.C., Rowden, A.A., Clark, M.R., Kilgallen, N.M., Linley, T., Mayor, D.J., Jamieson, A.J.** Community structure and diversity of scavenging amphipods from bathyal to hadal depths in three South Pacific Trenches. Deep-Sea Research I.– 2016. – V.111 – P. 121–137.
5. **Lang, W., Sirisansaneeyakul, S., Martins, L.O., Ngiwsara, L., Sakairi, N., Pathomaree, W., Okuyama, M., Mori, H., Kimura.** A Biodecolorization of a food azo dye by the deep sea Dermacoccus abyssi MT1.1T strain from the Mariana Trench. Journal of Environmental Management. – 2014. – V.132. – P. 155-164.
6. **Luo, M., Gieskes, J., Chen, L., Shi, X., Chen, D.** Provenances, distribution, and accumulation of organic matter in the southern Mariana Trench rim and slope: Implication for carbon cycle and burial in hadal trenches. Marine Geology. – 2017. – V.386, P.98–106.
7. **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 I, –2015. –V.100, P.21–33.