Plotting Ternary Diagrams by R Library ggtern for Geological Modelling

Polina Lemenkova

Plotting Ternary Diagrams by R Library *ggtern* for Geological Modelling

Polina LEMENKOVA[1*]

[1] *Ocean University of China, College of Marine Geo-sciences, Qingdao, China,*
pauline.lemenkova@gmail.com

Abstract

Selecting proper methods of data modelling is crucial in geosciences, as effective data visualization enables better understanding of complex geological phenomena: processes, structure and dynamics. Various approaches of the data analysis by R language include both traditional methods of linear charts and other approaches to data visualization: ternaries, circular and radar charts. Using ternaries for triple correlations between variable can be seen in applied geological analysis which proves it to be important method for data modelling.

Visualizing geological variables by ternaries enables to highlight correlation between variables in a triangle, how data are dependent and affected. In this study, several geologic, tectonic and geomorphic variables, such as sediment thickness, tectonic plates, volcanic areas, steepness and depths, were tested using R based modelling in {ggtern} library. Other graphs include radar charts and circular diagrams. Visualizing attributes as a triple component correlation by ternaries gives a better insight to the geological factors. Traditional techniques for visualization of pairwise linear correlations are not sufficient to show triple variations. Ternaries approach identifies data correlations by triple factors. Additional graphical models include circular and Euler-Venn diagrams of quantitative and qualitative geospatial data modelling. The study is supported by 7 R code listings and 9 figures.

Keywords: Ternary Diagrams, Geologic Modelling, R Programming, Data Analysis, Machine Learning.

Received: 26.04.2019
Revised: 20.11.2019
Accepted: 06.12.2019
[*]*Corresponding author:* Polina LEMENKOVA,
Ocean University of China, College of Marine Geo-sciences, Qingdao, China
E-mail: pauline.lemenkova@gmail.com
Cite this article as: P. Lemenkova, Plotting Ternary Diagrams by R Library *ggtern* for Geological Modelling, *Eastern Anatolian Journal of Science, Vol. 5, Issue 2, 16-25, 2019*

1. Introduction

1.1. Background

The scope of the current paper is visualization of the ternary models by means of the open source {ggtern} library of R programming developed and documented in described in (Hamilton, 2018). The paper aimed at testing and using library {ggtern}, an extension of ggplot2 R package (Wickham, 2009) in marine geological research for plotting correlations between geological factors: submarine sediment thickness, data distribution around volcanic zones and geomorphological parameters (slope steepness and aspect).

Ternary diagram is an important way of data visualization and modelling in applied geology and geosciences. Ternaries visualize a triple component correlation between the factors constituting a certain system representing them as triangular plots. Each side of the ternary diagrams corresponds to an individual factor or variable of the system.

Plotting ternaries is not as trivial as linear or curve plots showing dependancies between the variables. However, the majority of the research on the statistical modelling focus on more common representation of the data. Therefore, a certain attention should be given to the question of application of {ggtern} library to plot ternary diagrams for geological modelling.

1.2. Research Aim

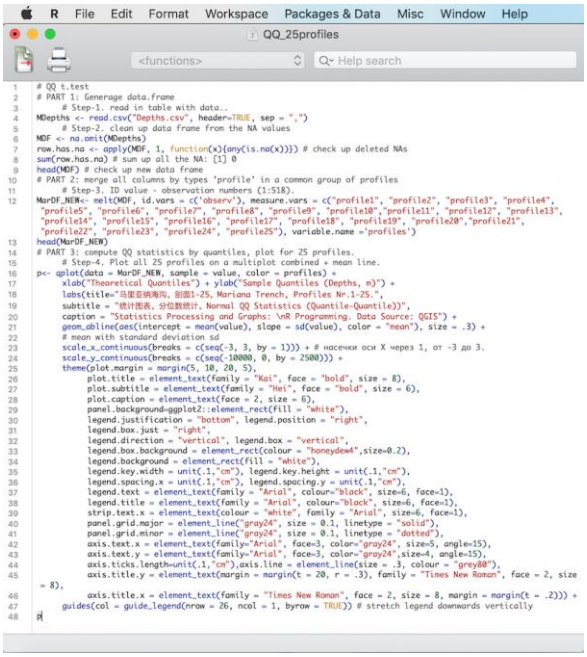
Understanding factors affecting geological structure of the oceanic trench are crucial for modelling geological variables for exploration resources. There are various approaches of data analysis aimed to study factors that may affect trench formation (e.g., Lemenkova, 2018b, 2019b).

A particular case study area of this research is Mariana Trench, and oceanic trench located in the west Pacific Ocean. The trench was cross-sectioned by the 25 profiles and in the points of the transecting the geological samples were recorded in a table. The table was processed by R libraries described below. Additionally, auxiliary plots were plotted by R

language libraries {fmsb}, {circlize}, {venn} and {ggplt2} visualizing radar charts, circular bar plots and Euler-Venn diagrams. All graphics aimed to show relationships in the geological variables.

1.3. Research Problem

Several studies report using ternary diagrams for visualizing triple correlations between the factors: (Fuhrman, Lindsley, 1988; Cosby et al., 1984; Markert etal., 2017). The importance of using ternaries in geology and soil studies depend on multi-variance dependancies of the factors that present the interplay of their correlation. For example, soil studies often use the percentages of the organic matter, clay and sand constituting soil structure. Intersection between the first and second variables and then adding the third variable depending on the first two show complex connections between the components of a geological or soil systems.




```
# QQ t.test
# PART 1: Generage data.frame
      # Step-1. read in table with data...
MDepths <- read.csv("Depths.csv", header=TRUE, sep = ",")
      # Step-2. clean up data frame from the NA values
MDF <- na.omit(MDepths)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))}) # check up deleted NAs
sum(row.has.na) # sum up all the NA: [1] 0
head(MDF) # check up new data frame
# PART 2: merge all columns by types 'profile' in a common group of profiles
      # Step-3. ID value - observation numbers (1:518).
MarDF_NEW<- melt(MDF, id.vars = c('observ'), measure.vars = c("profile1", "profile2", "profile3", "profile4",
 "profile5", "profile6", "profile7", "profile8", "profile9", "profile10","profile11", "profile12", "profile13",
 "profile14", "profile15", "profile16", "profile17", "profile18", "profile19", "profile20","profile21",
 "profile22", "profile23", "profile24", "profile25"), variable.name ='profiles')
head(MarDF_NEW)
# PART 3: compute QQ statistics by quantiles, plot for 25 profiles.
      # Step-4. Plot all 25 profiles on a multiplot combined + mean line.
p<- qplot(data = MarDF_NEW, sample = value, color = profiles) +
      xlab("Theoretical Quantiles") + ylab("Sample Quantiles (Depths, m)") +
      labs(title="马里亚纳海沟, 剖面1-25, Mariana Trench, Profiles Nr.1-25.",
      subtitle = "统计图表, 分位数统计, Normal QQ Statistics (Quantile-Quantile))",
      caption = "Statistics Processing and Graphs: \nR Programming. Data Source: QGIS") +
      geom_abline(aes(intercept = mean(value), slope = sd(value), color = "mean"), size = .3) +
      # mean with standard deviation sd
      scale_x_continuous(breaks = c(seq(-3, 3, by = 1))) + # насечки оси X через 1, от -3 до 3.
      scale_y_continuous(breaks = c(seq(-10000, 0, by = 2500))) +
      theme(plot.margin = margin(5, 10, 20, 5),
            plot.title = element_text(family = "Kai", face = "bold", size = 8),
            plot.subtitle = element_text(family = "Hei", face = "bold", size = 6),
            plot.caption = element_text(face = 2, size = 6),
            panel.background=ggplot2::element_rect(fill = "white"),
            legend.justification = "bottom", legend.position = "right",
            legend.box.just = "right",
            legend.direction = "vertical", legend.box = "vertical",
            legend.box.background = element_rect(colour = "honeydew4",size=0.2),
            legend.background = element_rect(fill = "white"),
            legend.key.width = unit(.1,"cm"), legend.key.height = unit(.1,"cm"),
            legend.spacing.x = unit(.1,"cm"), legend.spacing.y = unit(.1,"cm"),
            legend.text = element_text(family = "Arial", colour="black", size=6, face=1),
            legend.title = element_text(family = "Arial", colour="black", size=6, face=1),
            strip.text.x = element_text(colour = "white", family = "Arial", size=6, face=1),
            panel.grid.major = element_line("gray24", size = 0.1, linetype = "solid"),
            panel.grid.minor = element_line("gray24", size = 0.1, linetype = "dotted"),
            axis.text.x = element_text(family="Arial", face=3, color="gray24", size=5, angle=15),
            axis.text.y = element_text(family="Arial", face=3, color="gray24",size=4, angle=15),
            axis.ticks.length=unit(.1,"cm"),axis.line = element_line(size = .3, colour = "grey80"),
            axis.title.y = element_text(margin = margin(t = 20, r = .3), family = "Times New Roman", face = 2, size
= 8),
            axis.title.x = element_text(family = "Times New Roman", face = 2, size = 8, margin = margin(t = .2))) +
      guides(col = guide_legend(nrow = 26, ncol = 1, byrow = TRUE)) # stretch legend downwards vertically
#
```

Listing 1.

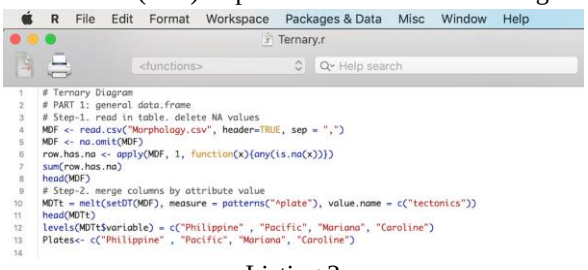
Due to the importance of the ternaries for the Earth sciences, graphical representation of ternaries has been developed in other statistical programs and languages, such as Matlab (Sandrock, 2016), Python (Harper, 2015; Rossum, 2011), and certain attempts for spread sheets integration were done by Excel (Marshall, 1996). However, the functionality provided by R language developed by (Wilkinson, 2005) significantly exceeds previous attempts both in terms of flexibility of data manipulation and graphical aesthetics (R Development Core Team, 2014).

The case study of the current data set is analysis of the marine geological factors of the Mariana Trench, a hadal deep-sea trench located in the west Pacific Ocean, that has a complex geomorphology and is well known for the deepest place on the Earth, the Challenger (Nakanishi and Hashimoto, 2011). A research aim of this study is to visualize the connections between multiple factors affecting the geomorphology of the trench: geology, geomorphology, bathymetry by means of ternaries and auxiliary plots (circular, radar charts, Euler-Venn diagrams showing logical connections).

2. Materials and Methods

The methodology includes using statistical algorithms by R programming language for plotting five types of the graphical visualization of geological data: ternary, that is the main aim of this research, and auxiliary plots, such as radar charts, circular plots and Euler-Venn logical diagrams.

The programming codes used for data modelling and graphical plotting are provided below. First step consists in computing a QQ (that is, quantile-quantile) plot. The QQ plot is a probability graph, which is a statistical method for comparing two probability distributions by visualizing their quantiles against each other (Figure 1). On the QQ plot we can see the depths values on Y axe and theoretical quantiles on the X axe (Figure 1). The offset between the line of the bathymetric measurements and the points of the samples shows that the mean of the sampling data varies. The R listing for the QQ plot (Figure 1) is as follows in 4 (four) steps 48 lines of R code: Listing 1.

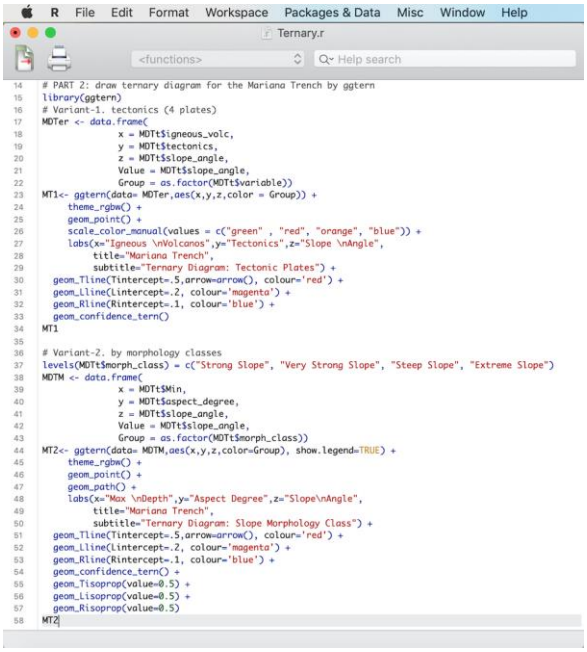



```
# Ternary Diagram
# PART 1: general data.frame
# Step-1. read in table. delete NA values
MDF <- read.csv("Morphology.csv", header=TRUE, sep = ",")
MDF <- na.omit(MDF)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))})
sum(row.has.na)
head(MDF)
# Step-2. merge columns by attribute value
MDTe = melt(setDT(MDF), measure = patterns("^plate"), value.name = c("tectonics"))
head(MDTe)
levels(MDTe$variable) = c("Philippine" , "Pacific", "Mariana", "Caroline")
Plates<- c("Philippine" , "Pacific", "Mariana", "Caroline")

```

Listing 2

The next step is the main part of the study, that consists in computing and visualizing triangular diagrams displaying the proportion of the three geological variables that sum together to a constant. Comparing to other statistical methods and

approaches, ternaries are notable for having a triangular coordinate system where the edges of the triangles are the axes showing variables.

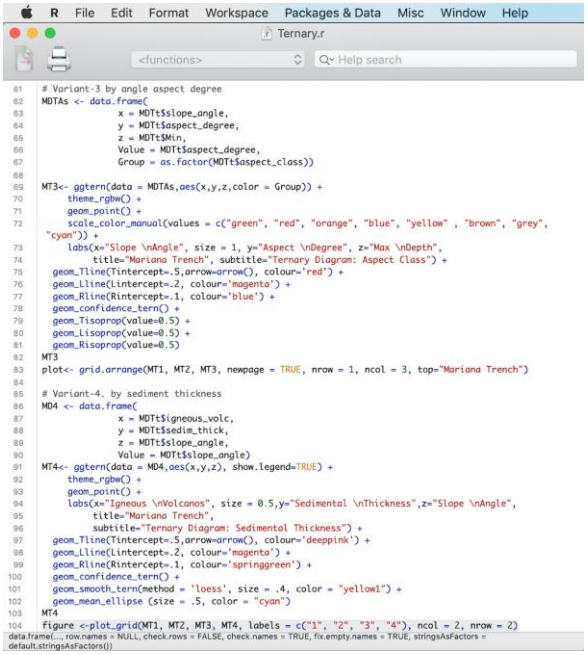


```
# PART 2: draw ternary diagram for the Mariana Trench by ggtern
library(ggtern)
# Variant-1. tectonics (4 plates)
MDTer <- data.frame(
                x = MDTt$igneous_volc,
                y = MDTt$tectonics,
                z = MDTt$slope_angle,
                Value = MDTt$slope_angle,
                Group = as.factor(MDTt$variable))
MT1<- ggtern(data= MDTer,aes(x,y,z,color = Group)) +
     theme_rgbw() +
     geom_point() +
     scale_color_manual(values = c("green" , "red", "orange", "blue")) +
     labs(x="Igneous \nVolcanos",y="Tectonics",z="Slope \nAngle",
          title="Mariana Trench",
          subtitle="Ternary Diagram: Tectonic Plates") +
  geom_Tline(Tintercept=.5,arrow=arrow(), colour='red') +
  geom_Lline(Lintercept=.2, colour='magenta') +
  geom_Rline(Rintercept=.1, colour='blue') +
  geom_confidence_tern()
MT1

# Variant-2. by morphology classes
levels(MDTt$morph_class) = c("Strong Slope", "Very Strong Slope", "Steep Slope", "Extreme Slope")
MDTM <- data.frame(
                x = MDTt$Min,
                y = MDTt$aspect_degree,
                z = MDTt$slope_angle,
                Value = MDTt$slope_angle,
                Group = as.factor(MDTt$morph_class))
MT2<- ggtern(data= MDTM,aes(x,y,z,color=Group), show.legend=TRUE) +
     theme_rgbw() +
     geom_point() +
     geom_path() +
     labs(x="Max \nDepth",y="Aspect Degree",z="Slope\nAngle",
          title="Mariana Trench",
          subtitle="Ternary Diagram: Slope Morphology Class") +
  geom_Tline(Tintercept=.5,arrow=arrow(), colour='red') +
  geom_Lline(Lintercept=.2, colour='magenta') +
  geom_Rline(Rintercept=.1, colour='blue') +
  geom_confidence_tern() +
  geom_Tisoprop(value=0.5) +
  geom_Lisoprop(value=0.5) +
  geom_Risoprop(value=0.5)
MT2
```

Listing 3.

The importance of ternaries consists in the approach of the simultaneous comparison of three variables, so that their impacts can be notable for changing conditions. Triangular or diagonal way of data visualization can be seen on correlation matrices (e.g., Lemenkova, 2018a; Lemenkova, 2019a).The programming R code for the ternary diagrams was modified after the package documented codes in 'ggtern' library (Hamilton 2018), an extension to ggplot2 (Wickham). The workflow is presented as follows in 3 (three) steps. Step 1 presented in Code listing 2 illustrates generating data frame and reading in initial data from the .csv table into the R environment.

Step 2 presented in Code listing 3 (lines 14-58 for the ternary diagrams code) produces two variants of the ternaries: by morphology classes and tectonic plates. Step 3 illustrates a case for Figure 3, that is, correlation between geological factors: 'Tectonic Plate', 'Igneous Volcanos' and 'Slopes Angle'. Step 3 was repeated for the plots illustrated on figure 2, 4, 5 and 6, respectively, by changing the names of the columns in the programming code as illustrated in Code listing 4. The ternaries are visualized in barycentric coordinates that represent triples of variables corresponding to the center masses placed at the vertices of a reference triangle (Bottema, 1982). Therefore, the dependence of the variables 1, 2 and 3 can be visually compared as a triple. The third part of the research consisted in plotting radar charts that are another statistical approach, different from the classic linear charts showing regression or correlation lines (Lemenkova, 2019d): circular diagrams.



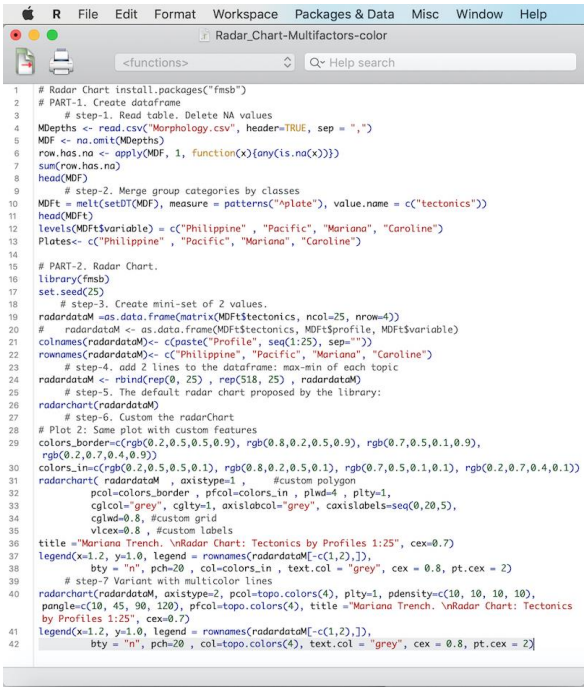
```
# Variant-3 by angle aspect degree
MDTAs <- data.frame(
                x = MDTt$slope_angle,
                y = MDTt$aspect_degree,
                z = MDTt$Min,
                Value = MDTt$aspect_degree,
                Group = as.factor(MDTt$aspect_class))

MT3<- ggtern(data = MDTAs,aes(x,y,z,color = Group)) +
     theme_rgbw() +
     geom_point() +
     scale_color_manual(values = c("green", "red", "orange", "blue", "yellow" , "brown", "grey",
"cyan")) +
     labs(x="Slope \nAngle", size = 1, y="Aspect \nDegree", z="Max \nDepth",
          title="Mariana Trench", subtitle="Ternary Diagram: Aspect Class") +
  geom_Tline(Tintercept=.5,arrow=arrow(), colour='red') +
  geom_Lline(Lintercept=.2, colour='magenta') +
  geom_Rline(Rintercept=.1, colour='blue') +
  geom_confidence_tern() +
  geom_Tisoprop(value=0.5) +
  geom_Lisoprop(value=0.5) +
  geom_Risoprop(value=0.5)
MT3
plot<- grid.arrange(MT1, MT2, MT3, newpage = TRUE, nrow = 1, ncol = 3, top="Mariana Trench")

# Variant-4. by sediment thickness
MD4 <- data.frame(
                x = MDTt$igneous_volc,
                y = MDTt$sedim_thick,
                z = MDTt$slope_angle,
                Value = MDTt$slope_angle)
MT4<- ggtern(data = MD4,aes(x,y,z), show.legend=TRUE) +
     theme_rgbw() +
     geom_point() +
     labs(x="Igneous \nVolcanos", size = 0.5,y="Sedimental \nThickness",z="Slope \nAngle",
          title="Mariana Trench",
          subtitle="Ternary Diagram: Sedimental Thickness") +
  geom_Tline(Tintercept=.5,arrow=arrow(), colour='deeppink') +
  geom_Lline(Lintercept=.2, colour='magenta') +
  geom_Rline(Rintercept=.1, colour='springgreen') +
  geom_confidence_tern() +
  geom_smooth_tern(method = 'loess', size = .4, color = "yellow1") +
  geom_mean_ellipse (size = .5, color = "cyan")
MT4
  figure <-plot_grid(MT1, MT2, MT3, MT4, labels = c("1", "2", "3", "4"), ncol = 2, nrow = 2)
data.frame(..., row.names = NULL, check.rows = FALSE, check.names = TRUE, fix.empty.names = TRUE, stringsAsFactors =
default.stringsAsFactors())
```

Listing 4.

Comparing to the traditional linear charts, radar charts represent a 2D chart of multivariate data comparison. Radar charts assign each variable an axis and plot the data as a multi-polygonal shape on the axes. In this study, radar chart shows the distribution of the samples by 25 bathymetric cross-section profiles. For example, we can see that profiles 5 and 6 have the least data, while profiles 19 to 24 are comparable (Figure 7). The radar chart (Figure 7) was plotted using R code by {fmsb} library in 7 (seven) steps 42 lines: Listing 5.

Radar charts is a useful approach to visualize a series of the multivariate quantitative observation samples. In such a case, each sample is represented by a polygon, so that one can see how the data overlap, if there are outliers among the variables and in which direction they correlate (Figure 7). Another approach to visualize data in non-traditional, non-linear way, is circular plot. Circular plots are visually round-shaped and show data distributed as strips with data points plotted along the unit circle. The length of each strip shows the value of the data (Figure 8). For

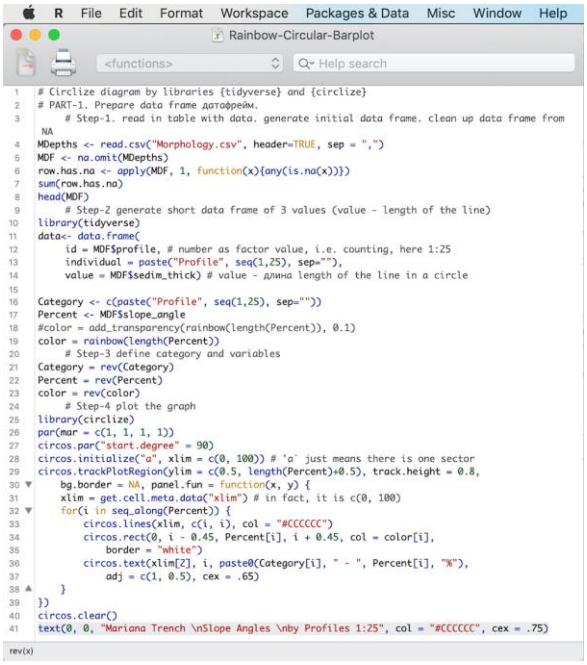
comparison, rainbow color palette was applied to better distinguish neighbor bins.




```
# Radar Chart install.packages("fmsb")
# PART-1. Create dataframe
    # step-1. Read table. Delete NA values
MDepths <- read.csv("Morphology.csv", header=TRUE, sep = ",")
MDF <- na.omit(MDepths)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))})
sum(row.has.na)
head(MDF)
    # step-2. Merge group categories by classes
MDFt = melt(setDT(MDF), measure = patterns("^plate"), value.name = c("tectonics"))
head(MDFt)
levels(MDFt$variable) = c("Philippine" , "Pacific", "Mariana", "Caroline")
Plates<- c("Philippine" , "Pacific", "Mariana", "Caroline")

# PART-2. Radar Chart.
library(fmsb)
set.seed(25)
    # step-3. Create mini-set of 2 values.
radardataM =as.data.frame(matrix(MDFt$tectonics, ncol=25, nrow=4))
#    radardataM <- as.data.frame(MDFt$tectonics, MDFt$profile, MDFt$variable)
colnames(radardataM)<- c(paste("Profile", seq(1:25), sep=""))
rownames(radardataM)<- c("Philippine", "Pacific", "Mariana", "Caroline")
    # step-4. add 2 lines to the dataframe: max-min of each topic
radardataM <- rbind(rep(0, 25) , rep(518, 25) , radardataM)
    # step-5. The default radar chart proposed by the library:
radarchart(radardataM)
    # step-6. Custom the radarChart
# Plot 2: Same plot with custom features
colors_border=c(rgb(0.2,0.5,0.5,0.9), rgb(0.8,0.2,0.5,0.9), rgb(0.7,0.5,0.1,0.9),
 rgb(0.2,0.7,0.4,0.9))
colors_in=c(rgb(0.2,0.5,0.5,0.1), rgb(0.8,0.2,0.5,0.1), rgb(0.7,0.5,0.1,0.1), rgb(0.2,0.7,0.4,0.1))
radarchart( radardataM  , axistype=1 ,     #custom polygon
          pcol=colors_border , pfcol=colors_in , plwd=4 , plty=1,
          cglcol="grey", cglty=1, axislabcol="grey", caxislabels=seq(0,20,5),
          cglwd=0.8, #custom grid
          vlcex=0.8 , #custom labels
title ="Mariana Trench. \nRadar Chart: Tectonics by Profiles 1:25", cex=0.7)
legend(x=1.2, y=1.0, legend = rownames(radardataM[-c(1,2),]),
          bty = "n", pch=20 , col=colors_in , text.col = "grey", cex = 0.8, pt.cex = 2)
    # step-7 Variant with multicolor lines
radarchart(radardataM, axistype=2, pcol=topo.colors(4), plty=1, pdensity=c(10, 10, 10, 10),
 pangle=c(10, 45, 90, 120), pfcol=topo.colors(4), title ="Mariana Trench. \nRadar Chart: Tectonics
 by Profiles 1:25", cex=0.7)
legend(x=1.2, y=1.0, legend = rownames(radardataM[-c(1,2),]),
          bty = "n", pch=20 , col=topo.colors(4), text.col = "grey", cex = 0.8, pt.cex = 2)
```

Listing 5.

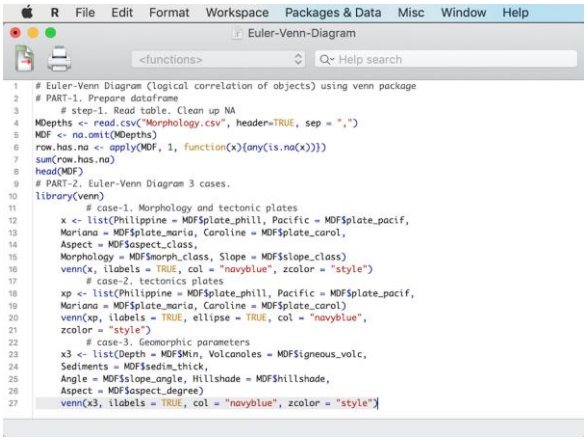
The range of the circle is divided into a 25 bins representing profiles (Figure 8).




```
# Circlize diagram by libraries {tidyverse} and {circlize}
# PART-1. Prepare data frame датафрейм.
    # Step-1. read in table with data. generate initial data frame. clean up data frame from
 NA
MDepths <- read.csv("Morphology.csv", header=TRUE, sep = ",")
MDF <- na.omit(MDepths)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))})
sum(row.has.na)
head(MDF)
    # Step-2 generate short data frame of 3 values (value - length of the line)
library(tidyverse)
data<- data.frame(
    id = MDF$profile, # number as factor value, i.e. counting, here 1:25
    individual = paste("Profile", seq(1,25), sep=""),
    value = MDF$sedim_thick) # value - длина length of the line in a circle

Category <- c(paste("Profile", seq(1,25), sep=""))
Percent <- MDF$slope_angle
#color = add_transparency(rainbow(length(Percent)), 0.1)
color = rainbow(length(Percent))
    # Step-3 define category and variables
Category = rev(Category)
Percent = rev(Percent)
color = rev(color)
    # Step-4 plot the graph
library(circlize)
par(mar = c(1, 1, 1, 1))
circos.par("start.degree" = 90)
circos.initialize("a", xlim = c(0, 100)) # 'a' just means there is one sector
circos.trackPlotRegion(ylim = c(0.5, length(Percent)+0.5), track.height = 0.8,
    bg.border = NA, panel.fun = function(x, y) {
    xlim = get.cell.meta.data("xlim") # in fact, it is c(0, 100)
    for(i in seq_along(Percent)) {
        circos.lines(xlim, c(i, i), col = "#CCCCCC")
        circos.rect(0, i - 0.45, Percent[i], i + 0.45, col = color[i],
            border = "white")
        circos.text(xlim[2], i, paste0(Category[i], " - ", Percent[i], "%"),
            adj = c(1, 0.5), cex = .65)
    }
})
circos.clear()
text(0, 0, "Mariana Trench \nSlope Angles \nby Profiles 1:25", col = "#CCCCCC", cex = .75)
```

Listing 6.

The length of each color strip shows the value: in this case, it is the degree slope gradient steepness changing by profiles. Thus, the samples are stacked in the bins of strips corresponding to the number of observations in each. The code for plotting circular diagram (Figure 8) was written by libraries {tidyverse} and {circlize} in 4 steps 41 lines, Listing 6. Final step of the data visualization consists in using Euler-Venn diagrams (Figure 9) as an approach to visualize logical relationships between the geologic




```
# Euler-Venn Diagram (logical correlation of objects) using venn package
# PART-1. Prepare dataframe
    # step-1. Read table. Clean up NA
MDepths <- read.csv("Morphology.csv", header=TRUE, sep = ",")
MDF <- na.omit(MDepths)
row.has.na <- apply(MDF, 1, function(x){any(is.na(x))})
sum(row.has.na)
head(MDF)
# PART-2. Euler-Venn Diagram 3 cases.
library(venn)
        # case-1. Morphology and tectonic plates
    x <- list(Philippine = MDF$plate_phill, Pacific = MDF$plate_pacif,
    Mariana = MDF$plate_maria, Caroline = MDF$plate_carol,
    Aspect = MDF$aspect_class,
    Morphology = MDF$morph_class, Slope = MDF$slope_class)
    venn(x, ilabels = TRUE, col = "navyblue", zcolor = "style")
        # case-2. tectonics plates
    xp <- list(Philippine = MDF$plate_phill, Pacific = MDF$plate_pacif,
    Mariana = MDF$plate_maria, Caroline = MDF$plate_carol)
    venn(xp, ilabels = TRUE, ellipse = TRUE, col = "navyblue",
    zcolor = "style")
        # case-3. Geomorphic parameters
    x3 <- list(Depth = MDF$Min, Volcanoles = MDF$igneous_volc,
    Sediments = MDF$sedim_thick,
    Angle = MDF$slope_angle, Hillshade = MDF$hillshade,
    Aspect = MDF$aspect_degree)
    venn(x3, ilabels = TRUE, col = "navyblue", zcolor = "style")
```

variables.

Listing 7.

The Euler-Venn diagrams were selected as an additional method of data visualization in this study, because, unlike linear statistical charts, they are suitable for highlighting complex hierarchies and overlapping relationships between the variables. The R listing used for plotting Euler-Venn diagram (Figure 10) was written using libraries in 2 steps, 3 cases 29 lines of the R code listing 7.

3. Results

The conceptual ideas of the statistical analysis for data science were derived from he existing works (Myers, and Well, 2003; Cielen et al., 2012) applied by R language.

3.1. Quantiles (QQ) plots

The Quantile-Quantile (QQ) plots (Figure 1) were plotted to visualize that the data can be approximated by a statistical distribution. The QQ is normally distributed. the horizontal and vertical axes of a QQ-plot visualize quantiles: X axe show theoretical quantiles, while Y axe show samples quantiles (showing depths in meters). The depths were derived from the observation points across 25 cross-section profiles digitized in a GIS project.

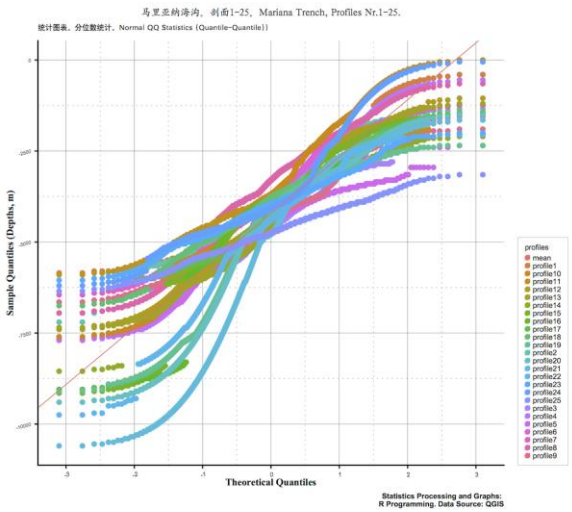


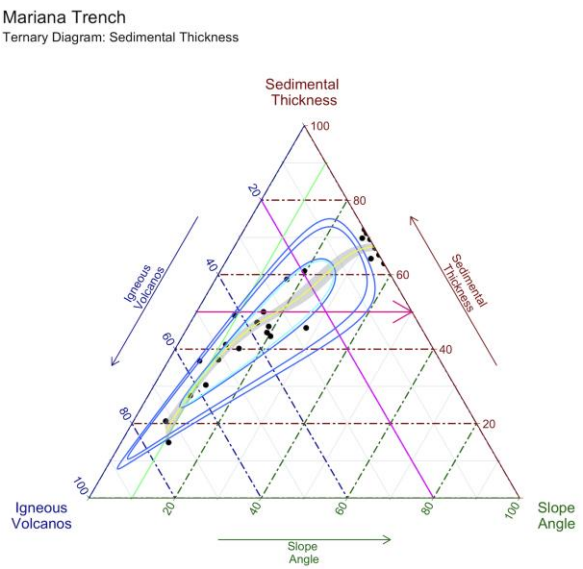
Figure 1.

The profiles cross Mariana Trench, an ocean hadal trench with complex geological and geomorphic settings. The QQ-plot shows (Figure 1) that the bathymetric data distribution (depth ranges) conform well enough to the normal distribution. Figure 1 is displaying probability of the normally distribution of a dataset of the observation samples of 25 cross-section bathymetric profiles.

The deviation between the theoretical quantiles and the normal distribution slightly increase in the lower left tail of the normal distribution, as shown on the graph. The discrepancy is also noticeable in the right tail of the normal distribution shown on Figure 1.

3.2. Ternary diagrams for geological modelling

Ternary diagrams displaying the correlation between sediment thickness, igneous volcanic areas and slope angles are shown on Figure 2. Here the dependance between the sediment thickness and volcanic activities in the study area can be seen as notable correlation. Figure 3 is displaying correlation between the geomorphic variables such as slope angles, sediment thickness and igneous volcanic areas (closeness of their location in meters towards the trench axis). Ternary diagrams were plotted (Figures 2-6) to show the triple correlation of several geological and geomorphic factors factors across four tectonic plates.



The observed factors include slope morphology classes, tectonics, igneous volcanic areas, aspect degree, slope angle, sediment thickness.

Figure 2.

Three dimensions of the ternary the diagram demonstrate complex relationship between the variables of the geological data set (Figure 2), which gives deeper insight into the triple dependancies between the factors rather than lines as in the standard diagrams. Figure 2 shows that the increase of the sediment thickness (starting from 0 to 100%) in the given sample points goes along with increase of the location of the igneous volcanic zones.

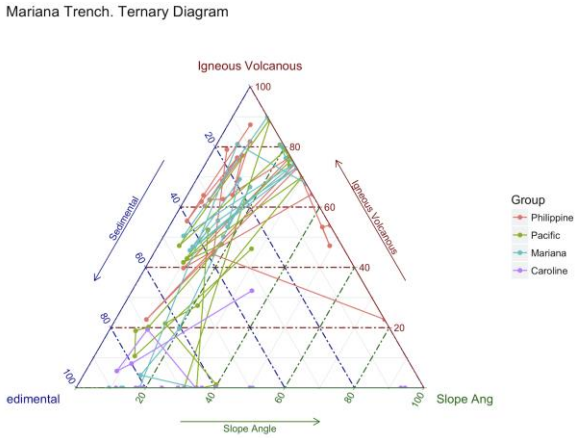


Figure 3.

Figure 3 shows the relationship of the three factors of the triangle constituting 100%. Geometry determines the shortest distances from the sampling point of the igneous volcanic zones, sediment thickness and slope angles to each of these three sides of the triangle, respectively.

Figure 4 is displaying correlation between the slope angles, aspect degree and maximal depths by observation samples in the Mariana Trench.

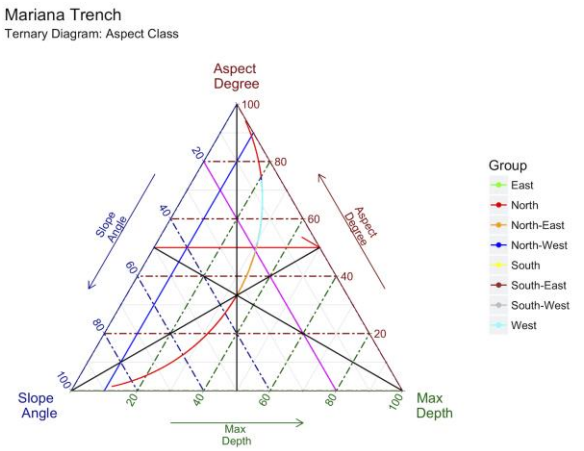


Figure 4.

Figure 5 is displaying correlation between the slope angles, maximal depths and aspect degree. In this case, the correlation between the volcanic activity and activation of the sedimentation rates is explained in the specialized reports on geology and volcanology (Chadwick et al., 2018; Contreras-Reyes et al., 2011).

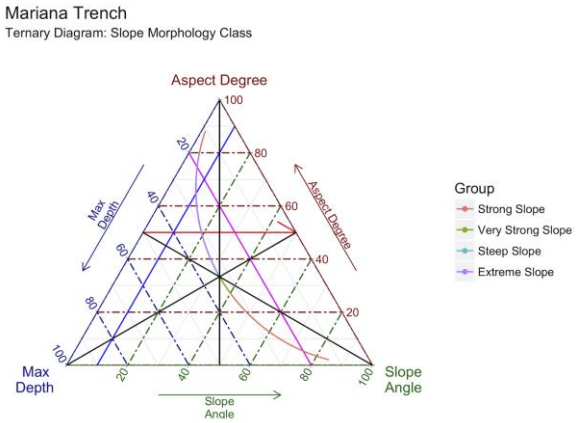


Figure 5.

Figure 4 and Figure 5 show dependancies between the geology and bathymetry, and between bathymetry and geomorphology, respectively. Figure 6 shows rather complex sample data distribution by four tectonic plates: Mariana, Caroline, Pacific and the Philippine Sea plate highlighted in colors. Tectonics and volcanism in the study area show the dependence between the variables while geomorphology, that is slope structure, is more affected by the bathymetric settings of the trench: depths in the observation samples. Figure 6 shows correlation between the slope angles, tectonics and igneous volcanic areas highlighting geologic and geomorphic settings.

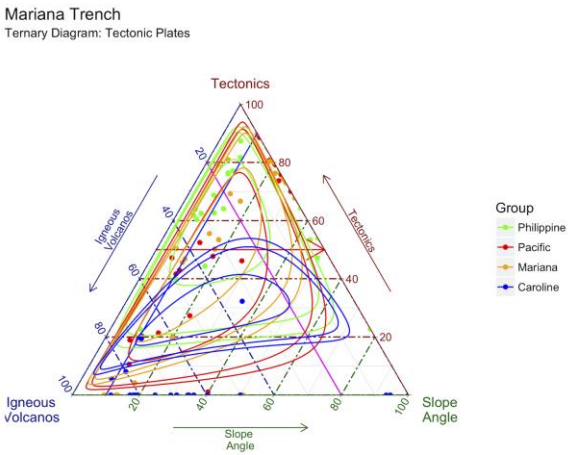


Figure 6.

The advantages of using a ternary plot for visualizing geological relationships is that three variables are conveniently plotted in a two-dimensional graph showing geological variables: sediment thickness, volcanism and slope steepness of the ocean trench. Ternaries shown on Figures 2 to 6 enabled to visualize correlation between the factors showing trips correlation between the categories.

3.3. Radar charts for data distribution

Radar chart (Figure 7) was visualized in this study with aim at detailed analysis of the tectonics by bathymetric profiles.

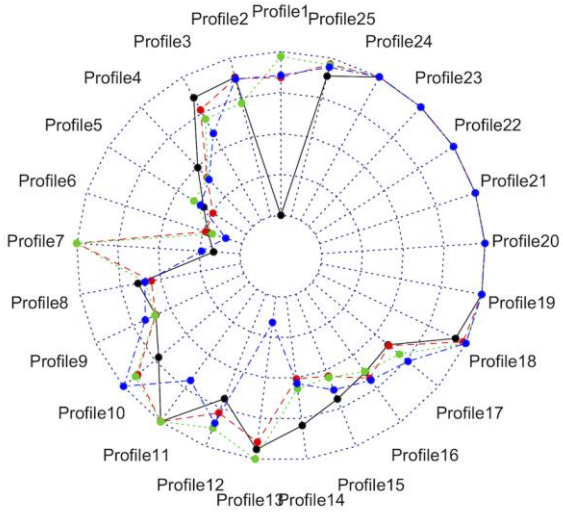


Figure 7.

Radar chart is displaying distribution of the observation samples by tectonic plates in a circular way. Line colors: ‘green’ – Mariana plate; ‘black’ – Philippine Sea plate; ‘blue’ – Caroline plate; ‘red’ – Pacific plate). Mariana Trench crosses four tectonic plates. The radar charts plotted using R library {fmsb} (Nakazawa, 2018), enabled to analyze data distribution by four tectonic plates.

As can be seen from the Figure 7, the cross-section profiles Nr. 5 and 6 are the shortest with only few observation points. Profile Nr. 7, crossing middle part of the trench, has sampling points mostly located on the Pacific tectonic plate while profiles 19-24 demonstrate homogeneous data distribution with

sample points mostly located on the Caroline plate (SW part). Radar chart enabled to visualize depth distribution by tectonic plates in a graphic representative way (Figure 7) and thus, to analyze geospatial parameters in relation with tectonic plates.

3.4. Circular bar plot for quantitative analysis

Circular plot (Figure 8) visualizes depth distribution and slope angle values using R code described in Methodology chapter derived from (Gu, 2018).

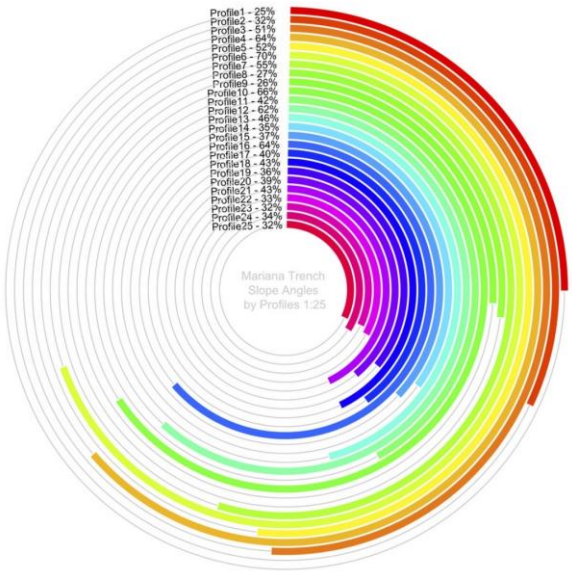


Figure 8.

The length of the stripe in a circle corresponds to the degree of the slope in a particular profile, that is, the most steep profiles have the longest strip on the circle. Here one can see the variations of the steepness in the slope degree by profiles, e.g. profiles Nr. 6 (lime green color) has the highest steepness, which is caused by the geomorphic structure of the slope. On the contrary, profiles Nr. 25-22 (located close to the inner circle, colored 'red' to 'magenta') have the least values of the slope angles. This visualization helps to analyze distribution of the geomorphological parameters by slope degree, which in turn, depends on the geological factors: underlying rock types and location close to the fault areas in the collision of the tectonic slabs.

3.5. Euler-Venn diagram for qualitative analysis

Euler-Venn diagram enables to comparing crossings between such categories as tectonic by four plates, slopes and geomorphology. Upon analysis of the possible combinations, intersections and similarities between the categories, the diagram was visualized as the relationship between the factors (Figure 9). Plotted Euler-Venn diagram derived algorithms and code developed by Duşa (2013). The area of the overlapping shapes is proportional to the number of crossing elements in a diagram, which is useful for explaining complex hierarchies and overlapping definitions (Thiem and Duşa, 2013).

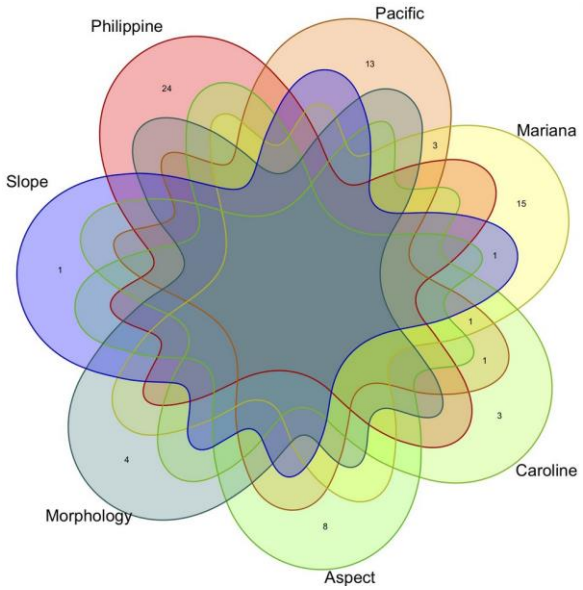


Figure 9.

With the case of factors affecting trench geomorphology, Euler-Venn diagram (Figure 9) was plotted by code given in the Methodology chapter. It shows logical relationships between variables of the data sets: tectonics, sediment thickness; geometric properties of the trench: slope angle, aspect class, hill shade group class derived as quantitative values from GIS; morphology; volcanism and depth values.

4. Conclusion

Nowadays, data processing by the machine learning algorithms is a crucial part of the geological information pool (Roberts, et al., 2018). The specific domain of the geology implies operating with big data due to the massive data sets taken as samples during geologic expeditions or observations recorded from the marine scientific cruises. Therefore, the value of the big data automatically modeled and processed by the machine is indisputable. Data analysis in marine and terrestrial geology can be performed through a variety of computer based approaches using existing

methodologies of the statistics and mathematics (Davis, 1990; Rossetier, 2017; Brownlee, 1965; Bulmer, 1979; Gauger et al., 2007; Lemenkova, 2019g; Kuhn et al., 2006). Processing large data sets by programming is a valuable approach of knowledge extraction in geosciences, due to the precision of the machine algorithms. Machine-generated geodata are a major resource for the Earth observation.

Current paper demonstrated data modelling and visualization by R programming aimed at visualizing correlations between the factors and variables with a particular focus on ternary plotting using available techniques (Hamilton, & Ferry, 2018; Mulcahy, 2012), as well as additional plotting of circular diagrams for modelling data distribution. Better understanding of the correlations and factors affecting structure and functioning of the geological bodies in general, and marine geological structure, such as ocean trenches, in particular, enables to take a deeper insights into the processes, geological drivers and environmental effects.

Besides R language, there are various approaches for the data analysis, modelling and visualization in marine geology, e.g. Python language (Lin, 2008; Lemenkova, 2019c), Gretl statistical package (Lemenkova, 2019e), SPSS Statistics (Lemenkova, 2019f), ArcGIS (Klaučo et al., 2013a, Klaučo et al., 2013b, Suetova et al., 2005a; Suetova et al., 2005b; Klaučo et al., 2014; Lemenkova et al., 2012; Klaučo et al., 2017), Autotrace digitizer for data processing aimed at bathymetric mapping (Schenke and Lemenkova, 2008). Among them, R language is notable for its functionality and reputability. A high-level programming language, R enables to perform all kind of the statistical data analysis, supported by a variety of algorithms in R packages. Effectiveness of R specifically consists in the advanced scientific statistical data analysis (Lemenkova, 2018c). Complex syntax with scripting approach flexibility of data processing (big tables, data frames). Using R language for data analysis in marine geology gives the power functionality to operate with large data sets.

As demonstrated in this paper, R language is an effective sophisticated tool enabling to operate with data frameworks created from the csv tables containing observation samples. It furthermore demonstrated the conceptual idea that rapidly developing new technologies of the programming languages and embedded statistical algorithms, with a case study of R libraries, can be successfully applied for the geological data modelling.

5. Acknowledgement

This research was funded by the China Scholarship Council, State Oceanic Administration, Marine Scholarship of China, Grant Nr. 2016SOA002, China.

References

BOTTEMA, O. (1982). On the Area of a Triangle in Barycentric Coordinates. *Crux. Math*. 8, 228–231.

BROWNLEE, K.A. (1979). *Statistical theory and methodology in science and engineering*. John Wiley & Sons, New York, 2nd edition.

BULMER, MG. *Principles of statistics*. Dover Publications, New York.

CHADWICK, W.W.Jr., MERLE, S.G., BAKER, E.T., WALKER, S.L., RESING, J.A., BUTTERFIELD, D.A., ANDERSON, M.O., BAUMBERGER, T., BOBBITT, A. M. (2018). A Recent Volcanic Eruption Discovered on the Central Mariana Back-Arc Spreading Center. *Frontiers in Earth Science*, 6, 1–16.

CIELEN, D., MEYSMAN, A.D.B., ALI, M. (2012). *Introducing Data Science. Big Data, Machine Learning and More, Using Python Tools*. Manning, Shelter Island, U.S.

CONTRERAS-REYES, E., CARRIZO, D. (2011). Control of high oceanic features and subduction channel on earthquake ruptures along the Chile-Peru subduction zone. *Physics of the Earth and Planetary Interiors*, 186, 49–58.

COSBY, B.J. HORNBERGER, G.M. CLAPP, R.B. GINN, T.R. (1984). A Statistical Exploration of the Relationships of Soil Moisture Characteristics to the Physical Properties of Soils. *Water Resources Research*, 20(6), 682–690. DOI: 10.1029/wr020i006p00682

DAVIS, J. (1990). *Statistics and Data Analysis in Geology*. Kansas Geological Survey John Wiley and Sons.

DUȘA, A. Venn package 1.7 version, R. https://www.rdocumentation.org/packages/venn/versions/1.7

ELKINS, L.T. GROVE, T.L. (1990). Ternary Feldspar Experiments and Thermodynamic Models. *American Mineralogist*, 75(5–6), 544–559.

FUHRMAN, M.L. LINDSLEY, D.H. (1988). Ternary-Feldspar Modeling and Thermometry. *American Mineralogist*, 73(3–4), 201–215.

GAUGER, S., KUHN, G., GOHL, K., FEIGL, T., LEMENKOVA, P., HILLENBRAND, C. (2007) Swath-bathymetric mapping. *The expedition ANTARKTIS-XXIII/4 of the Research Vessel 'Polarstern' in 2006. Reports on Polar and Marine Research*, 557, 38–45. DOI: 10.6084/m9.figshare.7439231

GU, Z. (2018). Package 'circlize' https://github.com/jokergoo/circlize

HAMILTON, N. (2018). ggtern: An Extension to ggplot2, for the Creation of Ternary Diagrams. R package v. 3.1.0: https://CRAN.R-project.org/package=ggtern

HAMILTON, N.E., FERRY, M. (2018). ggtern: Ternary Diagrams Using ggplot2. *Journal of Statistical Software, Code Snippets*, 87 (3), 1-17. DOI: 10.18637/jss.v087.c03

HARPER, M. (2015). python-ternary: Ternary Plots in Python. Zenodo. doi: 10.5281/zenodo.34938

KLAUČO, M., GREGOROVÁ, B., STANKOV, U., MARKOVIĆ, V., LEMENKOVA, P. (2013a). Determination of ecological significance based on geostatistical assessment: a case study from the Slovak Natura 2000 protected area. *Central European Journal of Geosciences*, 5(1), 28–42. DOI: 10.2478/s13533-012-0120-0

KLAUČO, M., GREGOROVÁ, B., STANKOV, U., MARKOVIĆ, V., LEMENKOVA, P. (2013b). Interpretation of Landscape Values, Typology and Quality Using Methods of Spatial Metrics for Ecological Planning. 54th Conference Environmental and Climate Technologies. Riga. DOI: 10.13140/RG.2.2.23026.96963

KLAUČO, M., GREGOROVÁ, B., STANKOV, U., MARKOVIĆ, V., LEMENKOVA, P. (2014). Landscape metrics as indicator for ecological significance: assessment of Sitno Natura 2000 sites, Slovakia. *Ecology and Environmental Protection. Proceedings of the International Conference*, 85–90. March 19–20, 2014. Minsk, DOI: 10.6084/m9.figshare.7434200

KLAUČO, M., GREGOROVÁ, B., STANKOV, U., MARKOVIĆ, V., LEMENKOVA, P. (2017). Land planning as a support for sustainable development based on tourism: A case study of Slovak Rural Region. *Environmental Engineering and Management Journal*, 2(16), 449-458. DOI: 10.30638/eemj.2017.045

KUHN, G., HASS, C., KOBER, M., PETITAT, M., FEIGL, T., HILLENBRAND, C. D., KRUGER, S., FORWICK, M., GAUGER, S., LEMENKOVA, P. (2006). The response of quaternary climatic cycles in the South-East Pacific: development of the opal belt and dynamics behavior of the West Antarctic ice sheet. In: Gohl, K. (ed). Expeditionsprogramm Nr. 75 ANT XXIII/4, AWI for Polar and Marine Research, Germany. DOI: 10.13140/RG.2.2.11468.87687

LEMENKOVA, P., PROMPER, C., GLADE, T. (2012). Economic Assessment of Landslide Risk for the Waidhofen a.d. Ybbs Region, Alpine Foreland, Lower Austria. *11th International Symposium on Landslides & the 2nd North American Symposium on Landslides & Engineered Slopes (NASL). Protecting Society through Improved Understanding*, Banff, Canada. 279–285. DOI: 10.6084/m9.figshare.7434230

LEMENKOVA, P. (2018a). 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, DOI: 10.6084/m9.figshare.7434167

LEMENKOVA, P. (2018b). 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, DOI: 10.6084/m9.figshare.7358207

LEMENKOVA, P. (2018c). Hierarchical Cluster Analysis by R language for Pattern Recognition in the Bathymetric Data Frame: a Case Study of the Mariana Trench, Pacific Ocean. *Virtual Simulation, Prototyping and Industrial Design*. 2(5), 147–152. doi: 10.6084/m9.figshare.7531550

LEMENKOVA, P. (2019a). Scatterplot Matrices of the Geomorphic Structure of the Mariana Trench at Four Tectonic Plates (Pacific, Philippine, Mariana and Caroline): a Geostatistical Analysis by R. *Problems of Tectonics of Continents and Oceans. 51st Tectonics Meeting*, 1, 347–352. Moscow: doi: 10.6084/m9.figshare.7699787

LEMENKOVA, P. (2019b). An Empirical Study of R Applications for Data Analysis in Marine Geology. *Marine Science and Technology Bulletin*, 8(1), 1–9. DOI: 10.33714/masteb.486678

LEMENKOVA, P. (2019c). Processing oceanographic data by Python libraries NumPy, SciPy and Pandas. Aquatic Research, vol. 2, pp. 73–91, doi: 10.3153/AR19009

LEMENKOVA, P. (2019d). Testing Linear Regressions by StatsModel Library of Python for Oceanological Data Interpretation. *Aquatic*

Sciences and Engineering, 34, 51–60, DOI: 10.26650/ASE2019547010

LEMENKOVA, P. (2019e). Regression Models by Gretl and R Statistical Packages for Data Analysis in Marine Geology. *International Journal of Environmental Trends*, 3(1), 39–59, doi: 10.6084/m9.figshare.8313362

LEMENKOVA, P. (2019f). Numerical Data Modelling and Classification in Marine Geology by the SPSS Statistics. *International Journal of Engineering Technologies*, 5(2), 90–99. doi: 10.6084/m9.figshare.8796941

LEMENKOVA, P. (2019g) Topographic surface modelling using raster grid datasets by GMT: example of the Kuril-Kamchatka Trench, Pacific Ocean. *Reports on Geodesy and Geoinformatics*, 108, 9-22. DOI: 10.2478/rgg-2019-0008

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. doi: 10.5194/gmd-2-1-2009

MARKERT, A. BOHNE, K. FACKLAM, M. WESSOLEK, G. (2017). Pedotransfer Functions of Soil Thermal Conductivity for the Textural Classes Sand, Silt, and Loam, *Soil Science Society of America Journal*, 81 (6), 1315–1327. DOI: 10.2136/sssaj2017.02.0062

MARSHALL, D. (1996). Ternplot: An Excel Spreadsheet for Ternary Diagrams. *Computers and Geosciences*, 22(6), 697–699. DOI: 10.1016/0098-3004(96)00012-x

MULCAHY, S.R. (2012). *Ternary Plots in R*. http://srmulcahy.github.io/2012/12/04/ternary-plots-r.html

MYERS, J.L. WELL, A.D. (2003). *Research Design and Statistical Analysis*. Ed. 2, Lawrence Erlbaum, U.S.

NAKANISHI, M., HASHIMOTO, J. (2011). A precise bathymetric map of the world's deepest seafloor, Challenger Deep in the Mariana Trench. *Marine Geophysical Researches*, 32(4), 455–463.

NAKAZAWA, M.R. *Documentation for ggradar, fmsb v0.6.3*, https://www.rdocumentation.org/packages/ggradar/versions/0.1/topics/ggradar

R DEVELOPMENT CORE TEAM. (2014). R: a language and environment for statistical computing. R Foundation for Statistical Computing URL: http://www.R-project.org Vienna, Austria.

ROSSUM, G. van. (2011). Python Programming Language. URL https://www.python.org/

ROBERTS, N.M., TIKOFF, B., DAVIS, J.R., STETSON-LEE, T. (2018). The utility of statistical analysis in structural geology. *Journal of Structural Geology*, 125, 1-39. doi: 10.1016/j.jsg.2018.05.030

ROSSETIER, D.G. (2017). *Tutorial: An example of statistical data analysis using the R environment for statistical computing*.

SANDROCK, C. (2016). *ternplot: Plots Ternary Phase Data on a Ternary Phase Diagram*. MATLAB File Exchange, 1.1, 2016-06-25. https://au.mathworks.com/matlabcentral/fileexchange/2299-alchemyst-ternplot

SCHENKE, H. W., LEMENKOVA, P. (2008). Zur Frage der Meeresboden-Kartographie: Die Nutzung von AutoTrace Digitizer für die Vektorisierung der Bathymetrischen Daten in der Petschora-See. *Hydrographische Nachrichten*, 25(81), 16–21. doi: 10.6084/m9.figshare.7435538

SUETOVA, I. A., USHAKOVA, L. A., & LEMENKOVA, P. (2005a). Geoinformation mapping of the Barents and Pechora Seas. Geography and Natural Resources, 4, 138–142. doi: 10.6084/m9.figshare.7435535

SUETOVA, I., USHAKOVA, L., LEMENKOVA, P. (2005b). Geoecological Mapping of the Barents Sea Using GIS. In: Digital Cartography & GIS for Sustainable Development of Territories. International Cartographic Conference ICC. La Coruña, España. DOI: 10.6084/m9.figshare.7435529

THIEM, A., DUȘA, A. (2013). *Introduction to R.* https://link.springer.com/chapter/10.1007%2F978-1-4614-4584-5_2

THIEM, A., DUȘA, A. (2013). *Qualitative comparative analysis with R: a user's guide*. Springer.

WICKHAM, H. (2009). g*gplot2: Elegant Graphics for Data Analysis*. Springer. http://ggplot2.org/

WILKINSON, L. (2005). *The Grammar of Graphics*. Statistics and Computing, 2nd edition. Springer.