İSTANBUL UNIVERSITY PRESS

EISSN 2602-473X

AQUATIC SCIENCES AND ENGINEERING

Aquat Sci Eng 2019; 34(2): 51-60 • DOI: https://doi.org/10.26650/ASE2019547010

Research Article

Testing Linear Regressions by StatsModel Library of Python for Oceanological Data Interpretation

Polina Lemenkova[1*]

Cite this article as: Lemenkova, P. (2019). Testing Linear Regressions by StatsModel Library of Python for Oceanological Data Interpretation. *Aquatic Sciences and Engineering, 34*(2), 51-60.

ORCID IDs of the authors:
P.L. 0000-0002-5759-1089

[1]Ocean University of China,
College of Marine Geo-sciences.
238 Songling Road, Laoshan,
266100, Qingdao, Shandong
Province, People's Republic of
China

Submitted:
30.03.2019

Revision Requested
22.05.2019

Last Revision Received
13.06.2019

Accepted:
13.06.2019

Online published:
26.06.2019

Correspondence:
Polina Lemenkova
E-mail:
lemenkovapolina@stu.ouc.edu.cn
pauline.lemenkova@gmail.com

ABSTRACT

The study area is focused on the Mariana Trench, west Pacific Ocean. The research aim is to investigate correlation between various factors, such as bathymetric depths, geomorphic shape, geographic location on four tectonic plates of the sampling points along the trench, and their influence on the geologic sediment thickness. Technically, the advantages of applying Python programming language for oceanographic data sets were tested. The methodological approaches include GIS data collecting, data analysis, statistical modelling, plotting and visualizing. Statistical methods include several algorithms that were tested: 1) weighted least square linear regression between geological variables, 2) autocorrelation; 3) design matrix, 4) ordinary least square regression, 5) quantile regression. The spatial and statistical analysis of the correlation of these factors aimed at the understanding, which geological and geodetic factors affect the distribution of the steepness and shape of the trench. Following factors were analysed: geology (sediment thickness), geographic location of the trench on four tectonics plates: Philippines, Pacific, Mariana and Caroline and bathymetry along the profiles: maximal and mean, minimal values, as well as the statistical calculations of the 1st and 3rd quantiles. The study revealed correlations between the sediment thickness and distinct variations of the trench geomorphology and sampling locations across various segments along the crescent of the trench.

Keywords: Programming language, Python, Statistical analysis, Pacific Ocean, Hadal trenches, Mariana Trench, oceanology, marine geology

INTRODUCTION

Multiple approaches and GIS methods have been used so far to model ocean seafloor, the most unreachable part of the Earth. These include echo sounding (Smith, & Sandwell, 1997), CTD (conductivity-temperature-depth profiler) technique (Taira et al., 2005), acoustic methods, continual profiling with single-beam systems and bottom coverage capability, multi-beam swath-mapping systems (Dierssen, & Theberge, 2014), classic approaches of the GIS mapping and other tools of geoinformatics (Fujie et al., 2006), remote sensing images analysis, navigation charts and data modelling using schematic cross-sections of the subduction zones (e.g. Schellart, 2008), statistical modelling using R and packages, e.g. dplyr, ggplot2, PMCMR, car (Reid et al., 2018). Of all these, statistical modelling of the oceanological data sets by means of R and Python programming languages is the most cost-effective for investigating hadal trench geomorphology.

Various studies have been reported on the geologic variations of the Mariana Trench involving uneven distribution of various geomorphic phenomena across the seafloor (e.g., Michibayashi et al., 2007; Grand et al., 1997). Amongst these, the questions of how the trench shape is varies and what are the factors affective its geomorphology are the most challenging in view of the importance of the deep ocean segments for the whole ocean environment. The distribution of elevations on the Earth or hypsography is highly uneven. Thus, the majority of the depths is occupied by deep basins (4– 6.5 km) while relatively few areas are covered by shallow zones. At the same time, a considerable pool of resources is hidden by the ocean depths which explains the actuality of the ocean research for the national economies. The limitations in marine geological methods are imposed by the high cost of the actual cruise marine expeditions. Using available open source geodata sets have removed this problem by the use of low-cost geospatial data and their processing in GIS and open source programming language R and Python. Similarly, Python based statistical set of libraries, such as NumPy, SciPy, StastModels, and Matplotlib statistical package present effective low-cost and easily available method for the marine oceanological data processing and modelling.

Regional studies of the marine geology of the trenches across the Pacific Ocean (e.g., Bello-González et al., 2018; Boston et al., 2017), modelling and predictions made upon analysis of the geophysical settings of various trench, produced by these investigators were instrumental in understanding current issues of the marine geological studies. The concepts of these reports on seafloor spreading, tectonic slab subduction, continental drift, and plate tectonics in the Pacific Ocean were analysed in the current research.

STUDY AREA AND DATA

The study area is located in the Mariana Trench, west Pacific Ocean, where the deepest place of the Earth is recorded (Theberge, 2008).

The geomorphology of the Mariana Trench was studied through the spatial and statistical analysis of the 25 cross-section bathymetric profiles digitized across the trench. Each profile has a length of 1000 km and a distance between each two is 100 km. The methodology consists of two parts: geospatial data processing and statistical analysis.

First, during the geospatial part of the research, the data were collected from the Quantum GIS project as vector layers. The attribute tables contained numerical data on bathymetry, geology, tectonic plates and geometric features of the Mariana Trench in its various segments of the geographic location: north-west, centre, south-west.

Second, during the statistical part of the research, the table in .csv format was then read into the Python environment using Pandas package. The profiles were observed using methods of the statistical modelling performed by Python programming language. During the statistical testing and experiment, several existing approaches (Box, & Tiao, 1992; Timm, 2007; Oliphant, 2007; Oliphant, 2015; Lemenkova, 2019) provided by the Python StatsModel and Matplotlib libraries were used as the core algorithms described below.

METHODOLOGY

Design matrix and model fit summary by the Ordinary Least Squares

The variables of geologic interest were stored in the table consisting of 18 rows where numeric information describes geology, bathymetry, geodesy and tectonics of the Mariana Trench. To fit most of the models covered by StatsModels Python library, the design or regressor matrix was created using existing approaches (Everitt, 2002; Box, & Tiao, 1992; Millman & Aivazis, 2011). The

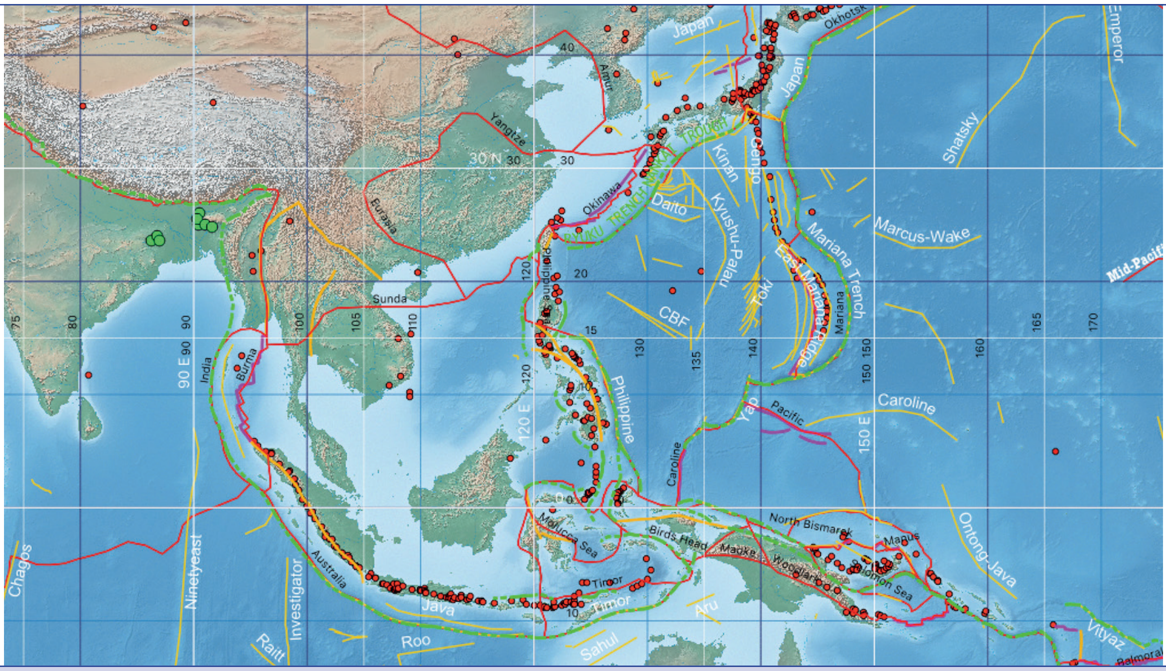
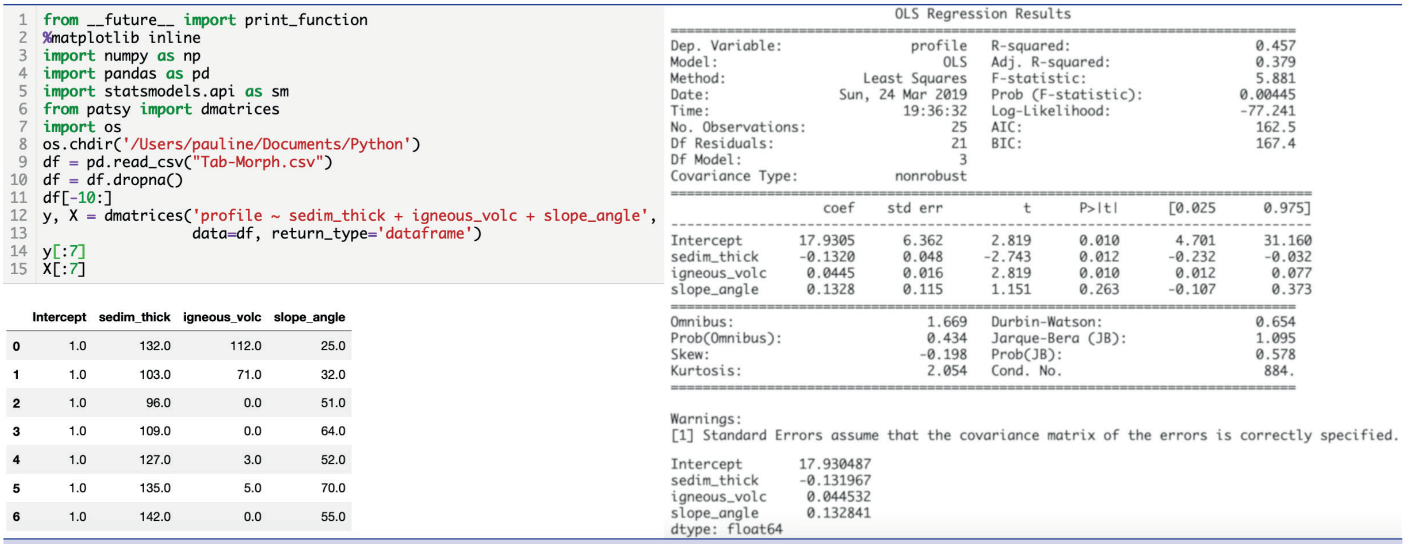


Figure 1. Study area: Mariana Trench, west Pacific Ocean (*Source: author, QGIS project*).



```
from __future__ import print_function
%matplotlib inline
import numpy as np
import pandas as pd
import statsmodels.api as sm
from patsy import dmatrices
import os
os.chdir('/Users/pauline/Documents/Python')
df = pd.read_csv("Tab-Morph.csv")
df = df.dropna()
df[-10:]
y, X = dmatrices('profile ~ sedim_thick + igneous_volc + slope_angle',
                 data=df, return_type='dataframe')
y[:7]
X[:7]
```

	Intercept	sedim_thick	igneous_volc	slope_angle
0	1.0	132.0	112.0	25.0
1	1.0	103.0	71.0	32.0
2	1.0	96.0	0.0	51.0
3	1.0	109.0	0.0	64.0
4	1.0	127.0	3.0	52.0
5	1.0	135.0	5.0	70.0
6	1.0	142.0	0.0	55.0

```
                            OLS Regression Results
==============================================================================
Dep. Variable:                profile   R-squared:                       0.457
Model:                            OLS   Adj. R-squared:                  0.379
Method:                 Least Squares   F-statistic:                     5.881
Date:                Sun, 24 Mar 2019   Prob (F-statistic):            0.00445
Time:                        19:36:32   Log-Likelihood:                -77.241
No. Observations:                  25   AIC:                             162.5
Df Residuals:                      21   BIC:                             167.4
Df Model:                           3
Covariance Type:            nonrobust
================================================================================
                   coef    std err          t      P>|t|      [0.025      0.975]
--------------------------------------------------------------------------------
Intercept       17.9305      6.362      2.819      0.010       4.701      31.160
sedim_thick     -0.1320      0.048     -2.743      0.012      -0.232      -0.032
igneous_volc     0.0445      0.016      2.819      0.010       0.012       0.077
slope_angle      0.1328      0.115      1.151      0.263      -0.107       0.373
==============================================================================
Omnibus:                        1.669   Durbin-Watson:                   0.654
Prob(Omnibus):                  0.434   Jarque-Bera (JB):                1.095
Skew:                          -0.198   Prob(JB):                        0.578
Kurtosis:                       2.054   Cond. No.                         884.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly specified.

Intercept       17.930487
sedim_thick     -0.131967
igneous_volc     0.044532
slope_angle      0.132841
dtype: float64
```

Table 1. Python code for design matrix 'dmatrices' (left) and OLS computation by StatsModels of Python (right), Mariana Trench data frame.

first is a matrix of endogenous variables of sediment thickness, which show the response or geological regressand on changed environmental conditions: geographic location, depth or tectonic plate.

The design matrix (Table 1, left) shows the results of the first six lines representing values of explanatory variables in a set of geological attributes, Mariana Trench.

The computing of the Ordinary Least Square (OLS) was based on the formula (1):

$$b_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})} \qquad (1)$$

Where,
n is the sample size;
x is a constant and a scalar regressor;
y is a random regressor, sampled together with *x*;
h is the number of lags being tested;

Each row of the calculated OLS coefficient estimates (Table 1, right) shows an individual bathymetric profile with the successive columns corresponding to the geologic and oceanographic variables and their specific values across the profiles.

Quantile statistics (QQ)

The used algorithm is very straightforward with a selected function of qqplot() by StatsModels to perform this task. The QQ regression is a common abbreviation for 'quantile by quantile' statistical plot. The plot shows (Figure 2) one quantile against another across various geological parameters (from left to right): A) Sediment thickness; B) Slope angle degrees; C) Pacific Plate; D); Philippine Plate E) Mariana Plate; F) Distribution of samples of igneous volcanic areas.

Technically, the plotting was performed using following code of Python for each corresponding plot:

```
ax1.plt = qqplot(df.sedim_thick, line='q', ax=ax1, fit=True,
linewidth=.5,alpha=.5,markerfacecolor='#00a497',
markeredgecolor='grey',)
```

The QQ statistics calculation has been based on the following formula (2) after Ljung, & Box (1978):

$$Q = n(n+2)\sum_{k=1}^{n} \frac{l^2_k}{(n-k)} \qquad (2)$$

Where,
n is the sample size;
rho is the sample autocorrelation at lag *k*, and
h is the number of lags being tested.

The comparison of all the six subplots enables to analyse the form of their shape against a straight line. The quantiles are bathymetric sample observations with geologic attribute values placed in the ascending order. The QQ statistics are used over the pool of the sampling data to study their distribution. A QQ statistic is a visual representation of the quantiles of a standard normal distribution of the geological data set across the Mariana Trench, showing their variation in space.

Weighted Least Squares

A Weighted Least Squares (WLS) for the geological variables are shown on Figure 3. The approach of a weighted least squares is a standard approach in regression analysis to approximate the solution of overdetermined systems which is the case for the complex marine geological systems. The least squares algorithms has two sub-types: linear or ordinary least squares and nonlinear least squares. In the scope of this research, only the linear least squares were tested: ordinary least squares and weight-

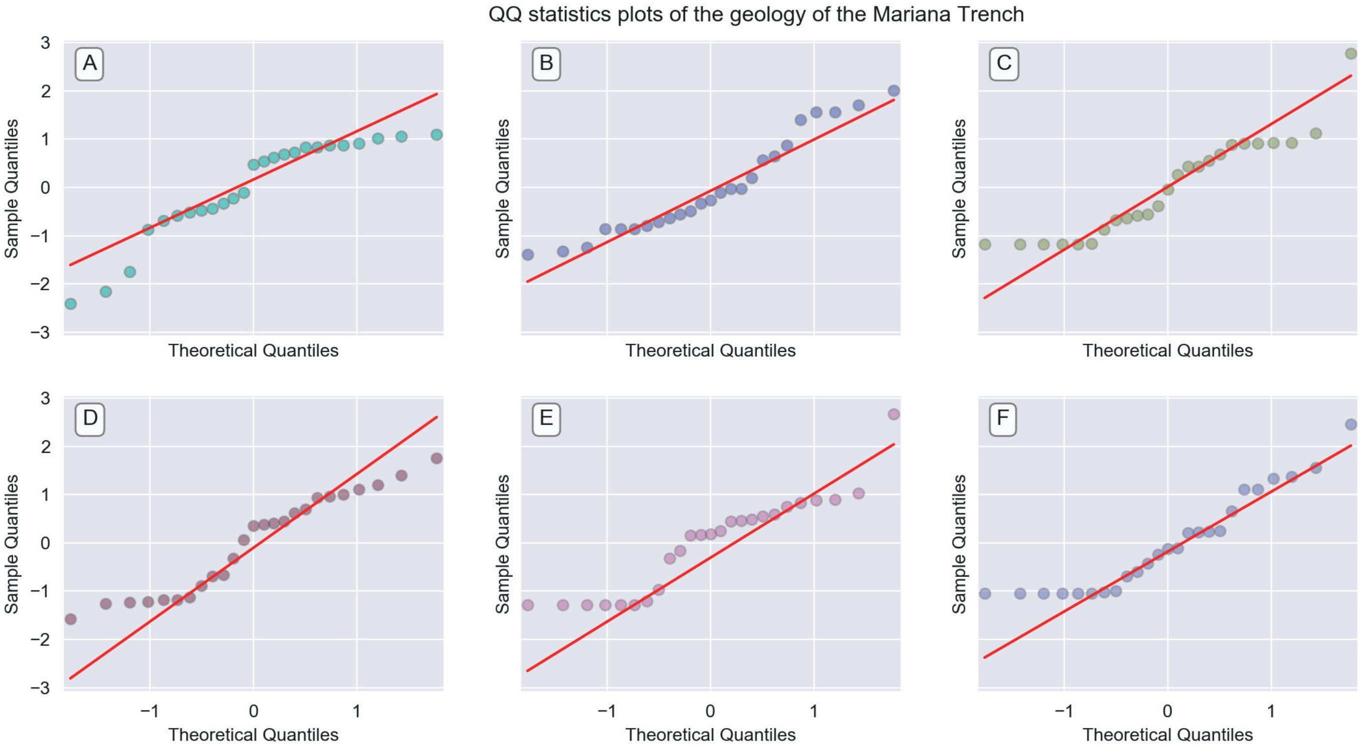


Figure 2. Plotted QQ statistics for the data distribution: A) Sediment thickness; B) Slope angle degrees; C) Pacific Plate; D); Philippine Plate E) Mariana Plate; F) Volcanic spots.

ed least square which is a generalization of the first one. In this case, the off-diagonal entries of the correlation matrix of the geological residuals are null the variances of the observations, yet unequal along the covariance matrix.

A WLS is a statistical approach representing a special case of the generalized least squares.

The calculation is based on the principle of the Gauss–Newton algorithm that solves a non-linear least squares problem by modifying a Newton's method for finding a minimum of a function. The computation was based on the general approach of existing equation (after Björck, 1996):

$$J^{T} WJ \Delta \beta = J^{T} W \Delta y \qquad (3)$$

Where
W is a diagonal when the observational errors are uncorrelated and the weight matrix;
J (t) is a transposed Jacobian matrix;
ß are unbiased estimators as linear column vectors, the entries of the Jacobian matrix;
y is a vector of the response values.

The calculations of the WLS for the data set (Table 2) were done according to the reported procedures (Seabold & Perktold, 2010; Strutz, T. (2016) by Python code snippet:

```
        # Step-1.
mod_wls = sm.WLS(y, X, weights=1./(w ** 2))
res_wls = mod_wls.fit()
print(res_wls.summary())
        # Step-2.
res_ols = sm.OLS(y, X).fit()
print(res_ols.params)
print(res_wls.params)
        # Step-3.
se  =  np.vstack([[res_wls.bse],  [res_ols.bse],
[res_ols.HC0_se],
                [res_ols.HC1_se], [res_ols.HC2_
se], [res_ols.HC3_se]])
se = np.round(se,4)
colnames = [‘x1’, ‘const’]
rownames = [‘WLS’, ‘OLS’, ‘OLS_HC0’, ‘OLS_HC1’,
‘OLS_HC3’, ‘OLS_HC3’]
tabl = SimpleTable(se, colnames, rownames, txt_
fmt=default_txt_fmt)
print(tabl)
```

Quantile regressions

Quantile regression shows (Figure 4) the estimated conditional median and other quantiles of the response geological variables. Thus the upper two rows of the plot show (Figure 4, A, B, C, D) data distribution across tectonic plates: Pacific Plate, Philippine Plate, Mariana Plate and Caroline Plate. The lower row of the plot (Figure 4, E, F) shows data distribution for the cumulative sediment thickness and slope angle degree by profiles.

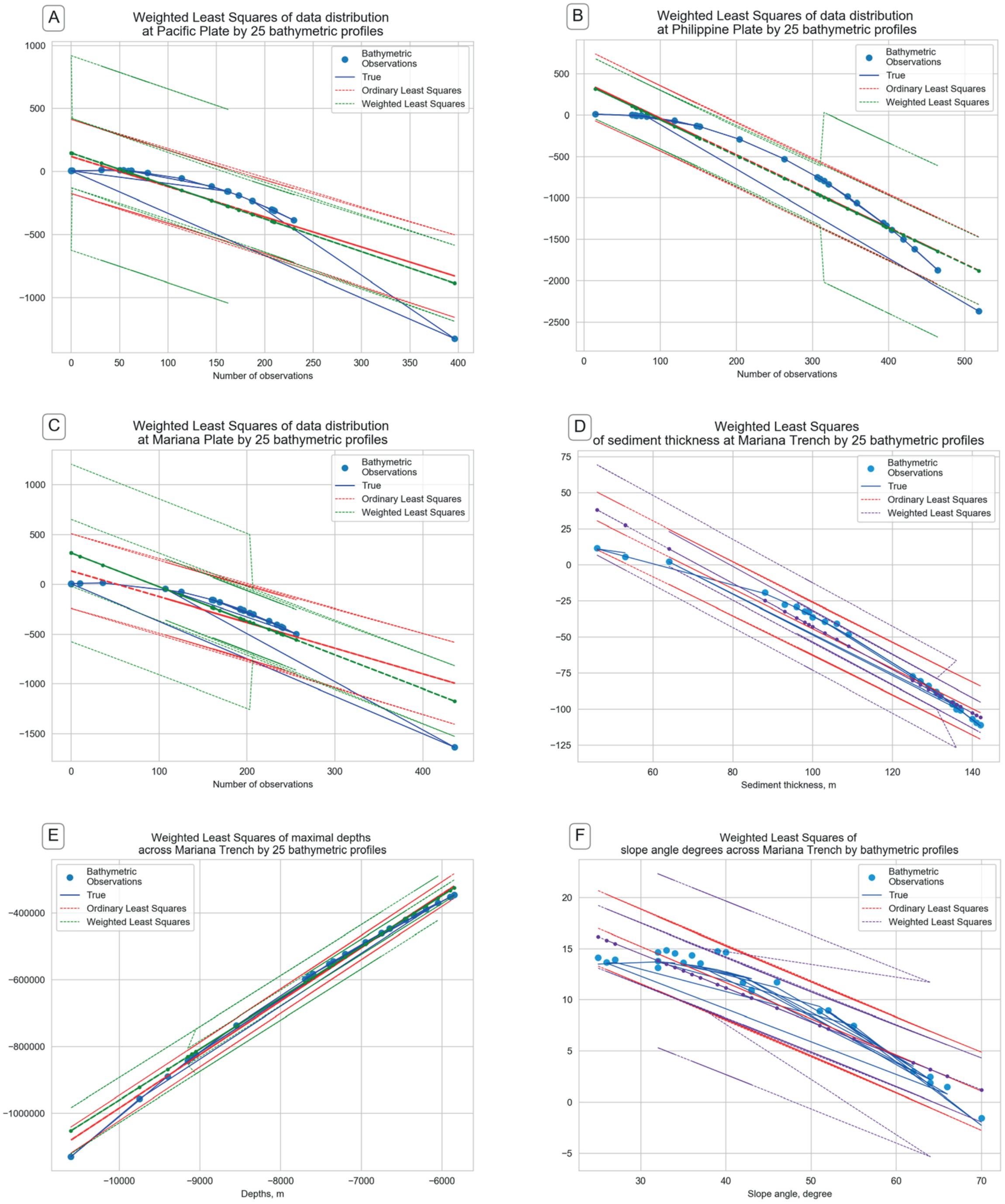
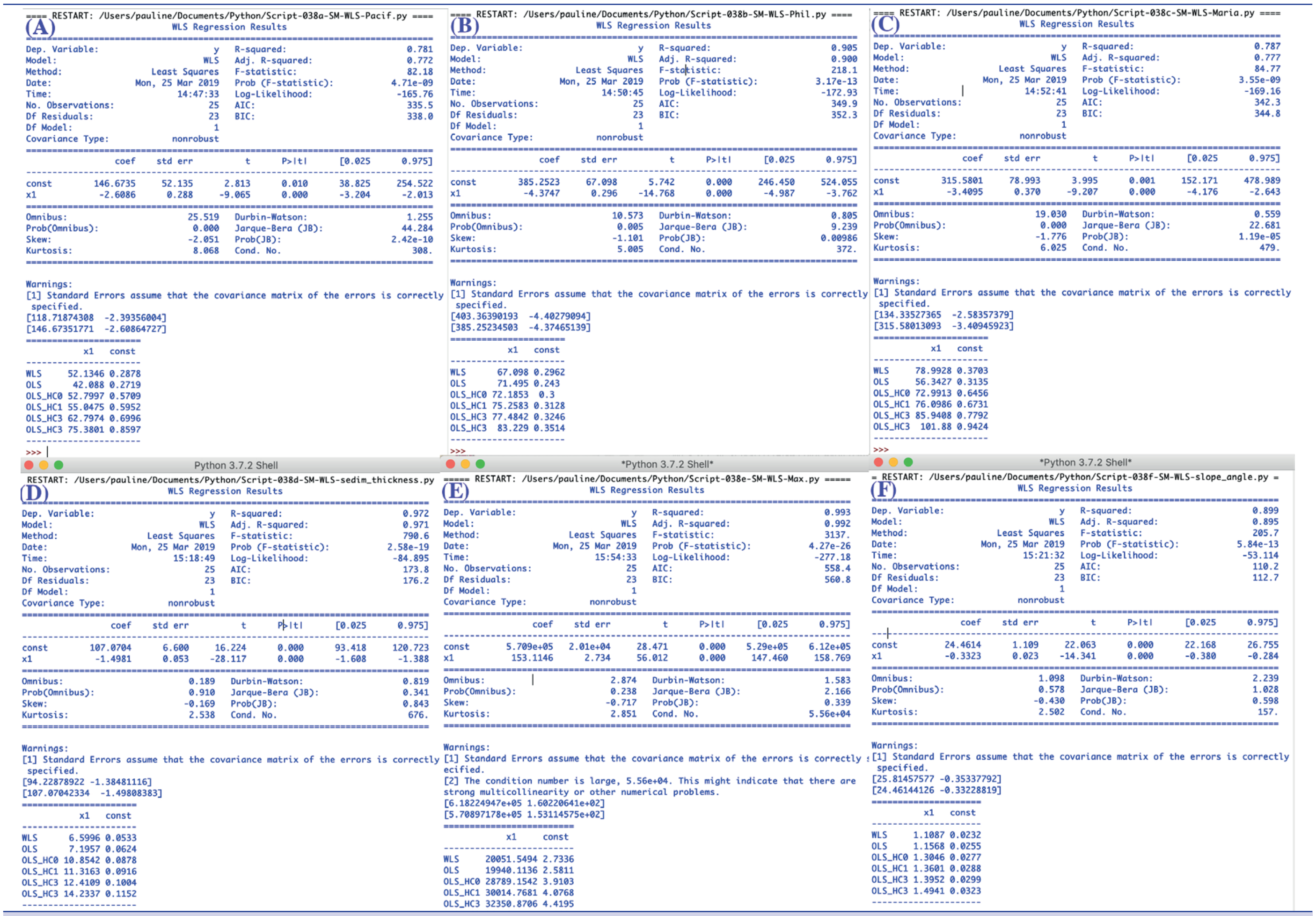


Figure 3. Weighted Least Squares plotted for data distribution: A) Pacific Plate, B) Philippine Plate, C) Mariana Plate, D) Sediment thickness, E) Depths (max); F) Slope angle degree.

The quantile regressions were plotted (Figure 4) using Python code by StatsModel:

```
    # Step-1. Least Absolute Deviation
mod = smf.quantreg('profile ~ slope_angle', data)
res = mod.fit(q=.5)
print(res.summary())
    # Step-2. Placing the quantile regression results in a Pandas DataFrame, and the OLS results in a dictionary
quantiles = np.arange(.05, .96, .1)
def fit_model(q):
    res = mod.fit(q=q)
      return [q, res.params['Intercept'], res.params['slope_angle']] + res.conf_int().loc['slope_angle'].tolist()
```



```
==== RESTART: /Users/pauline/Documents/Python/Script-038a-SM-WLS-Pacif.py ====
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.781
Model:                            WLS   Adj. R-squared:                  0.772
Method:                 Least Squares   F-statistic:                     82.18
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           4.71e-09
Time:                        14:47:33   Log-Likelihood:                -165.76
No. Observations:                  25   AIC:                             335.5
Df Residuals:                      23   BIC:                             338.0
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        146.6735     52.135      2.813      0.010      38.825     254.522
x1            -2.6086      0.288     -9.065      0.000      -3.204      -2.013
==============================================================================
Omnibus:                       25.519   Durbin-Watson:                   1.255
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               44.284
Skew:                          -2.051   Prob(JB):                     2.42e-10
Kurtosis:                       8.068   Cond. No.                         308.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
 specified.
[118.71874308  -2.39356004]
[146.67351771  -2.60864727]
======================
          x1   const
----------------------
WLS      52.1346 0.2878
OLS       42.088 0.2719
OLS_HC0 52.7997 0.5709
OLS_HC1 55.0475 0.5952
OLS_HC3 62.7974 0.6996
OLS_HC3 75.3801 0.8597
----------------------
>>> 
```

```
==== RESTART: /Users/pauline/Documents/Python/Script-038b-SM-WLS-Phil.py ====
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.905
Model:                            WLS   Adj. R-squared:                  0.900
Method:                 Least Squares   F-statistic:                     218.1
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           3.17e-13
Time:                        14:50:45   Log-Likelihood:                -172.93
No. Observations:                  25   AIC:                             349.9
Df Residuals:                      23   BIC:                             352.3
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        385.2523     67.098      5.742      0.000     246.450     524.055
x1            -4.3747      0.296    -14.768      0.000      -4.987      -3.762
==============================================================================
Omnibus:                       10.573   Durbin-Watson:                   0.805
Prob(Omnibus):                  0.005   Jarque-Bera (JB):                9.239
Skew:                          -1.101   Prob(JB):                      0.00986
Kurtosis:                       5.005   Cond. No.                         372.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
 specified.
[403.36390193  -4.40279094]
[385.25234503  -4.37465139]
======================
          x1   const
----------------------
WLS      67.098 0.2962
OLS      71.495 0.243
OLS_HC0 72.1853  0.3
OLS_HC1 75.2583 0.3128
OLS_HC3 77.4842 0.3246
OLS_HC3  83.229 0.3514
----------------------
>>>
```

```
==== RESTART: /Users/pauline/Documents/Python/Script-038c-SM-WLS-Maria.py ====
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.787
Model:                            WLS   Adj. R-squared:                  0.777
Method:                 Least Squares   F-statistic:                     84.77
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           3.55e-09
Time:                        14:52:41   Log-Likelihood:                -169.16
No. Observations:                  25   AIC:                             342.3
Df Residuals:                      23   BIC:                             344.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        315.5801     78.993      3.995      0.001     152.171     478.989
x1            -3.4095      0.370     -9.207      0.000      -4.176      -2.643
==============================================================================
Omnibus:                       19.030   Durbin-Watson:                   0.559
Prob(Omnibus):                  0.000   Jarque-Bera (JB):               22.681
Skew:                          -1.776   Prob(JB):                     1.19e-05
Kurtosis:                       6.025   Cond. No.                         479.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
 specified.
[134.33527365  -2.58357379]
[315.58013093  -3.40945923]
======================
          x1   const
----------------------
WLS      78.9928 0.3703
OLS      56.3427 0.3135
OLS_HC0 72.9913 0.6456
OLS_HC1 76.0986 0.6731
OLS_HC3 85.9408 0.7792
OLS_HC3  101.88 0.9424
----------------------
>>>
```

```
 RESTART: /Users/pauline/Documents/Python/Script-038d-SM-WLS-sedim_thickness.py
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.972
Model:                            WLS   Adj. R-squared:                  0.971
Method:                 Least Squares   F-statistic:                     790.6
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           2.58e-19
Time:                        15:18:49   Log-Likelihood:                -84.895
No. Observations:                  25   AIC:                             173.8
Df Residuals:                      23   BIC:                             176.2
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const        107.0704      6.600     16.224      0.000      93.418     120.723
x1            -1.4981      0.053    -28.117      0.000      -1.608      -1.388
==============================================================================
Omnibus:                        0.189   Durbin-Watson:                   0.819
Prob(Omnibus):                  0.910   Jarque-Bera (JB):                0.341
Skew:                          -0.169   Prob(JB):                        0.843
Kurtosis:                       2.538   Cond. No.                         676.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
 specified.
[94.22878922 -1.38481116]
[107.07042334  -1.49808383]
======================
          x1   const
----------------------
WLS       6.5996 0.0533
OLS       7.1957 0.0624
OLS_HC0 10.8542 0.0878
OLS_HC1 11.3163 0.0916
OLS_HC3 12.4109 0.1004
OLS_HC3 14.2337 0.1152
----------------------
```

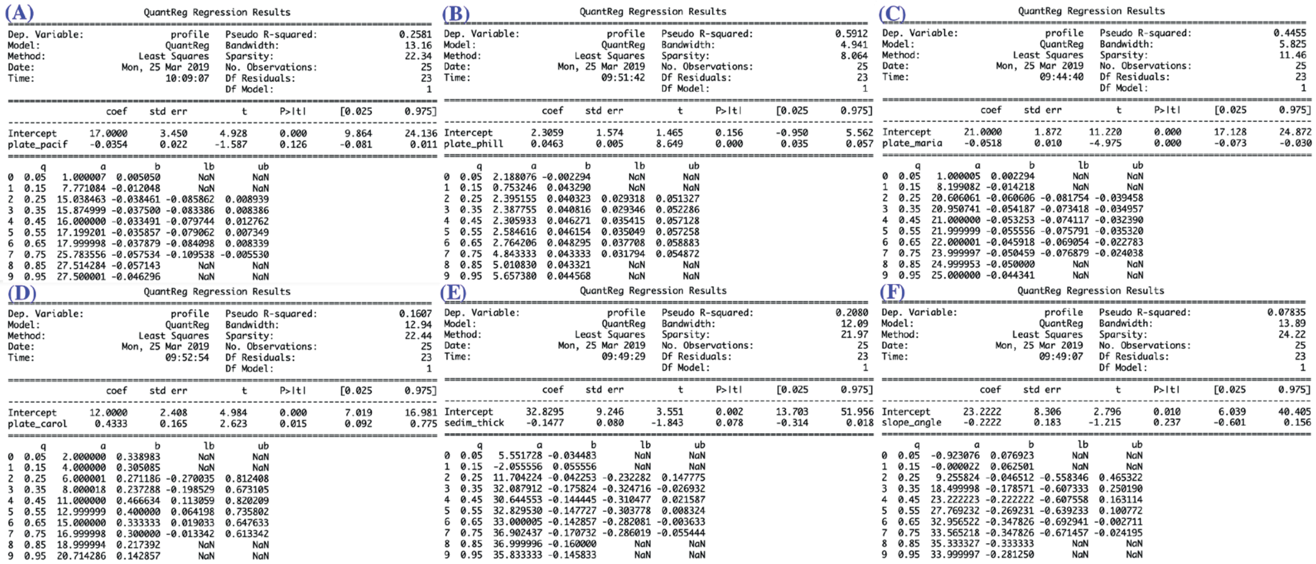
```
===== RESTART: /Users/pauline/Documents/Python/Script-038e-SM-WLS-Max.py =====
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.993
Model:                            WLS   Adj. R-squared:                  0.992
Method:                 Least Squares   F-statistic:                     3137.
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           4.27e-26
Time:                        15:54:33   Log-Likelihood:                -277.18
No. Observations:                  25   AIC:                             558.4
Df Residuals:                      23   BIC:                             560.8
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const       5.709e+05   2.01e+04     28.471      0.000    5.29e+05    6.12e+05
x1           153.1146      2.734     56.012      0.000     147.460     158.769
==============================================================================
Omnibus:                        2.874   Durbin-Watson:                   1.583
Prob(Omnibus):                  0.238   Jarque-Bera (JB):                2.166
Skew:                          -0.717   Prob(JB):                        0.339
Kurtosis:                       2.851   Cond. No.                     5.56e+04
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly s
ecified.
[2] The condition number is large, 5.56e+04. This might indicate that there are
strong multicollinearity or other numerical problems.
[6.18224947e+05 1.60220641e+02]
[5.70897178e+05 1.53114575e+02]
=========================
            x1     const
-------------------------
WLS     20051.5494 2.7336
OLS     19940.1136 2.5811
OLS_HC0 28789.1542 3.9103
OLS_HC1 30014.7681 4.0768
OLS_HC3 32350.8706 4.4195
```

```
= RESTART: /Users/pauline/Documents/Python/Script-038f-SM-WLS-slope_angle.py =
                            WLS Regression Results
==============================================================================
Dep. Variable:                      y   R-squared:                       0.899
Model:                            WLS   Adj. R-squared:                  0.895
Method:                 Least Squares   F-statistic:                     205.7
Date:                Mon, 25 Mar 2019   Prob (F-statistic):           5.84e-13
Time:                        15:21:32   Log-Likelihood:                -53.114
No. Observations:                  25   AIC:                             110.2
Df Residuals:                      23   BIC:                             112.7
Df Model:                           1
Covariance Type:            nonrobust
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
const         24.4614      1.109     22.063      0.000      22.168      26.755
x1            -0.3323      0.023    -14.341      0.000      -0.380      -0.284
==============================================================================
Omnibus:                        1.098   Durbin-Watson:                   2.239
Prob(Omnibus):                  0.578   Jarque-Bera (JB):                1.028
Skew:                          -0.430   Prob(JB):                        0.598
Kurtosis:                       2.502   Cond. No.                         157.
==============================================================================

Warnings:
[1] Standard Errors assume that the covariance matrix of the errors is correctly
 specified.
[25.81457577 -0.35337792]
[24.46144126 -0.33228819]
=====================
          x1   const
---------------------
WLS      1.1087 0.0232
OLS      1.1568 0.0255
OLS_HC0 1.3046 0.0277
OLS_HC1 1.3601 0.0288
OLS_HC3 1.3952 0.0299
OLS_HC3 1.4941 0.0323
---------------------
```

Table 2. Computed results of the WLS modelling for the data distribution by plates: A) Pacific; B) Philippine; C) Mariana; D) Sediment thickness; E) Depths; F) Slope angle.



```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:               0.2581
Model:                       QuantReg   Bandwidth:                       13.16
Method:                 Least Squares   Sparsity:                        22.34
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        10:09:07   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     17.0000      3.450      4.928      0.000       9.864      24.136
plate_pacif   -0.0354      0.022     -1.587      0.126      -0.081       0.011
==============================================================================
      q          a         b        lb        ub
0  0.05   1.000007  0.005050       NaN       NaN
1  0.15   7.771084 -0.012048       NaN       NaN
2  0.25  15.038463 -0.038461 -0.085862  0.008939
3  0.35  15.874999 -0.037500 -0.083386  0.008386
4  0.45  16.000000 -0.033491 -0.079744  0.012762
5  0.55  17.199201 -0.035857 -0.079062  0.007349
6  0.65  17.999998 -0.037879 -0.084098  0.008339
7  0.75  25.783556 -0.057534 -0.109538 -0.005530
8  0.85  27.514284 -0.057143       NaN       NaN
9  0.95  27.500001 -0.046296       NaN       NaN
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:               0.5912
Model:                       QuantReg   Bandwidth:                       4.941
Method:                 Least Squares   Sparsity:                        8.064
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        09:51:42   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept      2.3059      1.574      1.465      0.156      -0.950       5.562
plate_phill    0.0463      0.005      8.649      0.000       0.035       0.057
==============================================================================
      q         a         b        lb        ub
0  0.05  2.188076 -0.002294       NaN       NaN
1  0.15  0.753246  0.043290       NaN       NaN
2  0.25  2.395155  0.040323  0.029318  0.051327
3  0.35  2.387755  0.040816  0.029346  0.052286
4  0.45  2.305933  0.046271  0.035415  0.057128
5  0.55  2.584616  0.046154  0.035049  0.057258
6  0.65  2.764206  0.048295  0.037708  0.058883
7  0.75  4.843333  0.043333  0.031794  0.054872
8  0.85  5.010830  0.043321       NaN       NaN
9  0.95  5.657380  0.044568       NaN       NaN
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:               0.4455
Model:                       QuantReg   Bandwidth:                       5.825
Method:                 Least Squares   Sparsity:                        11.46
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        09:44:40   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     21.0000      1.872     11.220      0.000      17.128      24.872
plate_maria   -0.0518      0.010     -4.975      0.000      -0.073      -0.030
==============================================================================
      q          a         b        lb        ub
0  0.05   1.000005  0.002294       NaN       NaN
1  0.15   8.199082 -0.014218       NaN       NaN
2  0.25  20.606061 -0.060606 -0.081754 -0.039458
3  0.35  20.950741 -0.054187 -0.073418 -0.034957
4  0.45  21.000000 -0.053253 -0.074117 -0.032390
5  0.55  21.999999 -0.055556 -0.075791 -0.035320
6  0.65  22.000001 -0.045918 -0.069054 -0.022783
7  0.75  23.999997 -0.050459 -0.076879 -0.024038
8  0.85  24.999953 -0.050000       NaN       NaN
9  0.95  25.000000 -0.044341       NaN       NaN
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:               0.1607
Model:                       QuantReg   Bandwidth:                       12.94
Method:                 Least Squares   Sparsity:                        22.44
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        09:52:54   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     12.0000      2.408      4.984      0.000       7.019      16.981
plate_carol    0.4333      0.165      2.623      0.015       0.092       0.775
==============================================================================
      q          a         b        lb        ub
0  0.05   2.000000  0.338983       NaN       NaN
1  0.15   4.000000  0.305085       NaN       NaN
2  0.25   6.000001  0.271186 -0.270035  0.812408
3  0.35   8.000018  0.237288 -0.198529  0.673105
4  0.45  11.000000  0.466634  0.113059  0.820209
5  0.55  12.999999  0.400000  0.064198  0.735802
6  0.65  15.000000  0.333333  0.019033  0.647633
7  0.75  16.999998  0.300000 -0.013342  0.613342
8  0.85  18.999994  0.217392       NaN       NaN
9  0.95  20.714286  0.142857       NaN       NaN
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:               0.2080
Model:                       QuantReg   Bandwidth:                       12.09
Method:                 Least Squares   Sparsity:                        21.97
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        09:49:29   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     32.8295      9.246      3.551      0.002      13.703      51.956
sedim_thick   -0.1477      0.080     -1.843      0.078      -0.314       0.018
==============================================================================
      q          a         b        lb        ub
0  0.05   5.551728 -0.034483       NaN       NaN
1  0.15  -2.055556  0.055556       NaN       NaN
2  0.25  11.704224 -0.042253 -0.232282  0.147775
3  0.35  32.087912 -0.175824 -0.324716 -0.026932
4  0.45  30.644553 -0.144445 -0.310477  0.021587
5  0.55  32.829530 -0.147727 -0.303778  0.008324
6  0.65  33.000005 -0.142857 -0.282081 -0.003633
7  0.75  36.902437 -0.170732 -0.286019 -0.055444
8  0.85  36.999996 -0.160000       NaN       NaN
9  0.95  35.833333 -0.145833       NaN       NaN
```

```
                         QuantReg Regression Results
==============================================================================
Dep. Variable:                profile   Pseudo R-squared:              0.07835
Model:                       QuantReg   Bandwidth:                       13.89
Method:                 Least Squares   Sparsity:                        24.22
Date:                Mon, 25 Mar 2019   No. Observations:                   25
Time:                        09:49:07   Df Residuals:                       23
                                        Df Model:                            1
==============================================================================
                 coef    std err          t      P>|t|      [0.025      0.975]
------------------------------------------------------------------------------
Intercept     23.2222      8.306      2.796      0.010       6.039      40.405
slope_angle   -0.2222      0.183     -1.215      0.237      -0.601       0.156
==============================================================================
      q          a         b        lb        ub
0  0.05  -0.923076  0.076923       NaN       NaN
1  0.15  -0.000022  0.062501       NaN       NaN
2  0.25   9.255824 -0.046512 -0.558346  0.465322
3  0.35  18.499998 -0.178571 -0.607333  0.250190
4  0.45  23.222223 -0.222222 -0.607558  0.163114
5  0.55  27.769232 -0.269231 -0.639233  0.100772
6  0.65  32.956522 -0.347826 -0.692941 -0.002711
7  0.75  33.565218 -0.347826 -0.671457 -0.024195
8  0.85  35.333327 -0.333333       NaN       NaN
9  0.95  33.999997 -0.281250       NaN       NaN
```

Table 3. Results of the computations for the quantile regression for sediment thickness versus geologic parameters: A) Pacific Plate, B) Philippine Plate, C) Mariana Plate, D) Caroline Plate, E) Cumulative sediment thickness and F) Slope angle degree by profiles.

```
models = [fit_model(x) for x in quantiles]
models = pd.DataFrame(models, columns=[‘q’, ‘a’,
‘b’,’lb’,’ub’])
ols = smf.ols(‘profile ~ slope_angle’, data).fit()
ols_ci    =    ols.conf_int().loc[‘slope_angle’].
tolist()
```

```
ols = dict(a = ols.params[‘Intercept’],
           b = ols.params[‘slope_angle’],
           lb = ols_ci[0],
           ub = ols_ci[1])
print(models)
print(ols)
```

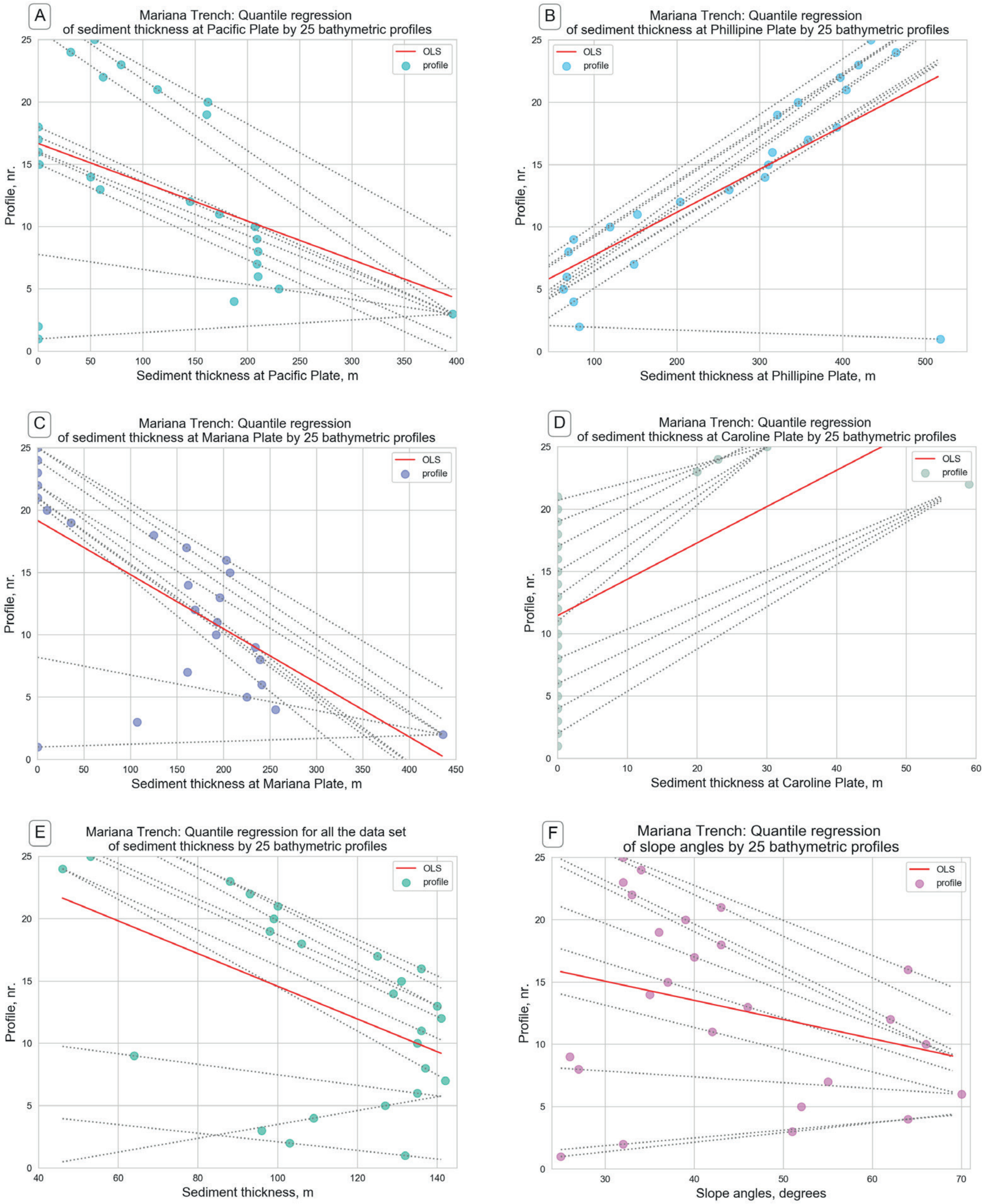


Figure 4. Quantile regression plotted for sediment thickness (m) versus geologic parameters: A) Pacific Plate, B) Philippine Plate, C) Mariana Plate, D) Caroline Plate, E) Cumulative sediment thickness and F) Slope angle degree by profiles.

Essentially, quantile regression is another approach of the linear regression tested in the current research. The quantitative results of the quantile regression are shown on Table 3, with respect to the relevant plots shown on Figure 4 (corresponding to Figure 4: A, B, C, D, E, F).

Dynamic regression model: State Autoregressive Moving Average (SARIMA)

The methodology of the dynamic regression model is based on the StatsModels embedded algorithm (Seabold & Perktold, 2010).

The abbreviation of SARIMA is the Space AutoRegressIve Moving Average model, initially developed by Ansley & Kohn, (1985). The concept of the application of the SARIMA time series estimation and post-estimation lies in the following. The changes in the variables that are taken as time lags in classic dynamic regression models are applied towards bathymetric profiles from 1 to 25. In this way, unlike in time series case when the SARIMA fits the univariate models with time-dependent disturbances, this case applies space-dependent disturbances, crossing the sampling selection from 1 to 25 (X axe).

Because the model includes both dependent and independent variables, the selected type of SARIMA was SARIMAX (see the Python code snippet below). The first group consists in changing variables that is geologic settings and bathymetry (depths). The second group (independent variables) is presented by the profiles lags that cross the Mariana Trench with the distance between each of 100 km and the length of 1000 km. This cross-section profiles are taken as independent variables. Therefore, the dependant variables differ spatially in different segments of the trench.

The model (Figure 5) fits univariate model of the geomorphic structure of the trench by independent values of the distribution of the bathymetric observation by profiles with dependent disturbances of depths.

Fitting the model was done using the Python snippet:

```
mod   =   sm.tsa.statespace.SARIMAX(data['sedim_
thick'], trend='c', order=(1,1,1))
res = mod.fit(disp=False)
print(res.summary())
fig   =   sm.graphics.tsa.plot_pacf(data.iloc[1:]
['Ddf.geology'], lags=25, ax=axes[1])
```

The algorithm fits a model where the disturbances follow a linear specification of the bathymetric distribution across the trench. The dependent and independent geological variables vary by profiles (Figure 5). Plotting was done using sublplot function of Matplotlib:

```
fig, axes = plt.subplots(2, 2, figsize=(15,8))
```

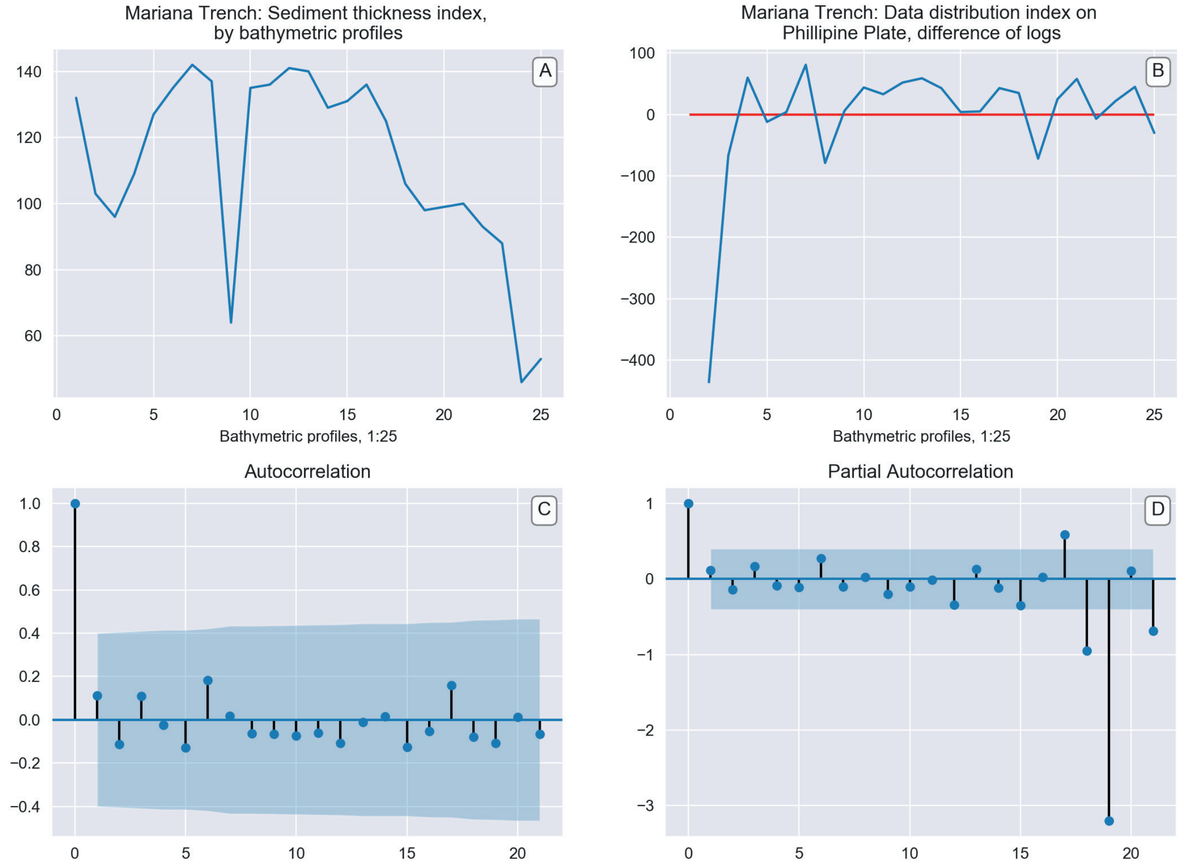
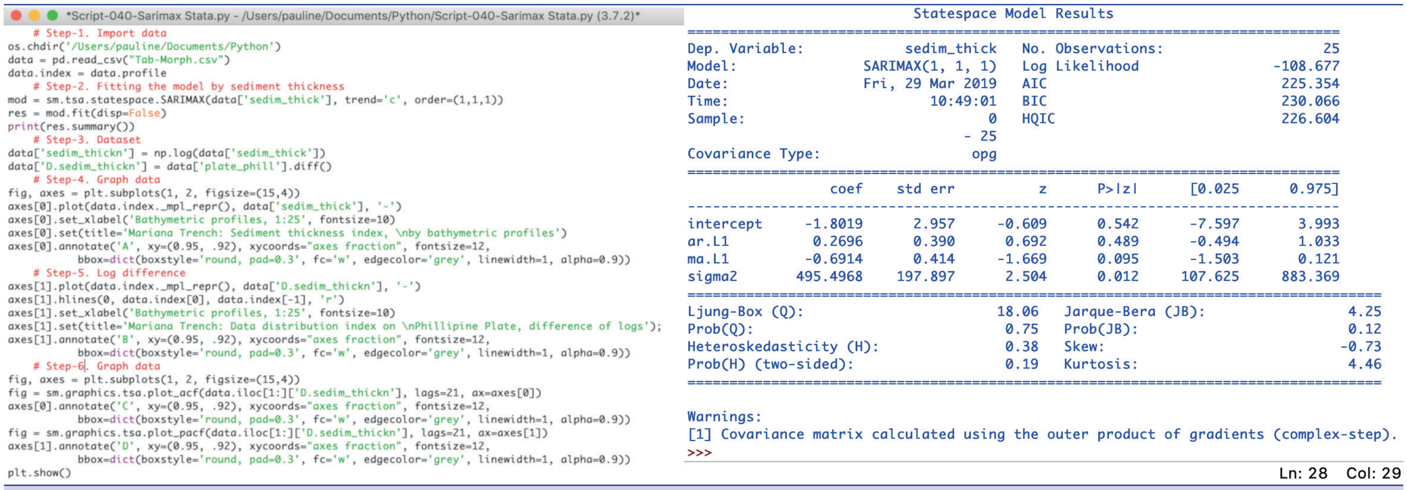


Figure 5. Plotted SARIMAX statistics for the bathymetry: A) Sediment thickness index; B) Data distribution index, Philippine Plate; C) Autocorrelation; D); Partial Autocorrelation.



Script-040-Sarimax Stata.py - /Users/pauline/Documents/Python/Script-040-Sarimax Stata.py (3.7.2)

```python
    # Step-1. Import data
os.chdir('/Users/pauline/Documents/Python')
data = pd.read_csv("Tab-Morph.csv")
data.index = data.profile
    # Step-2. Fitting the model by sediment thickness
mod = sm.tsa.statespace.SARIMAX(data['sedim_thick'], trend='c', order=(1,1,1))
res = mod.fit(disp=False)
print(res.summary())
    # Step-3. Dataset
data['sedim_thickn'] = np.log(data['sedim_thick'])
data['D.sedim_thickn'] = data['plate_phill'].diff()
    # Step-4. Graph data
fig, axes = plt.subplots(1, 2, figsize=(15,4))
axes[0].plot(data.index._mpl_repr(), data['sedim_thick'], '-')
axes[0].set_xlabel('Bathymetric profiles, 1:25', fontsize=10)
axes[0].set(title='Mariana Trench: Sediment thickness index, \nby bathymetric profiles')
axes[0].annotate('A', xy=(0.95, .92), xycoords="axes fraction", fontsize=12,
           bbox=dict(boxstyle='round, pad=0.3', fc='w', edgecolor='grey', linewidth=1, alpha=0.9))
    # Step-5. Log difference
axes[1].plot(data.index._mpl_repr(), data['D.sedim_thickn'], '-')
axes[1].hlines(0, data.index[0], data.index[-1], 'r')
axes[1].set_xlabel('Bathymetric profiles, 1:25', fontsize=10)
axes[1].set(title='Mariana Trench: Data distribution index on \nPhillipine Plate, difference of logs');
axes[1].annotate('B', xy=(0.95, .92), xycoords="axes fraction", fontsize=12,
           bbox=dict(boxstyle='round, pad=0.3', fc='w', edgecolor='grey', linewidth=1, alpha=0.9))
    # Step-6. Graph data
fig, axes = plt.subplots(1, 2, figsize=(15,4))
fig = sm.graphics.tsa.plot_acf(data.iloc[1:]['D.sedim_thickn'], lags=21, ax=axes[0])
axes[0].annotate('C', xy=(0.95, .92), xycoords="axes fraction", fontsize=12,
           bbox=dict(boxstyle='round, pad=0.3', fc='w', edgecolor='grey', linewidth=1, alpha=0.9))
fig = sm.graphics.tsa.plot_pacf(data.iloc[1:]['D.sedim_thickn'], lags=21, ax=axes[1])
axes[1].annotate('D', xy=(0.95, .92), xycoords="axes fraction", fontsize=12,
           bbox=dict(boxstyle='round, pad=0.3', fc='w', edgecolor='grey', linewidth=1, alpha=0.9))
plt.show()
```

```
                           Statespace Model Results
==============================================================================
Dep. Variable:            sedim_thick   No. Observations:                   25
Model:               SARIMAX(1, 1, 1)   Log Likelihood                -108.677
Date:                Fri, 29 Mar 2019   AIC                            225.354
Time:                        10:49:01   BIC                            230.066
Sample:                             0   HQIC                           226.604
                                 - 25
Covariance Type:                  opg
==============================================================================
                 coef    std err          z      P>|z|      [0.025      0.975]
------------------------------------------------------------------------------
intercept     -1.8019      2.957     -0.609      0.542      -7.597       3.993
ar.L1          0.2696      0.390      0.692      0.489      -0.494       1.033
ma.L1         -0.6914      0.414     -1.669      0.095      -1.503       0.121
sigma2       495.4968    197.897      2.504      0.012     107.625     883.369
===================================================================================
Ljung-Box (Q):                       18.06   Jarque-Bera (JB):                 4.25
Prob(Q):                              0.75   Prob(JB):                         0.12
Heteroskedasticity (H):               0.38   Skew:                            -0.73
Prob(H) (two-sided):                  0.19   Kurtosis:                         4.46
===================================================================================

Warnings:
[1] Covariance matrix calculated using the outer product of gradients (complex-step).
>>>
```

Table 4. Python programming code for used SARIMAX algorithm (left) and Statespace Model Results (right). Tested case in Mariana Trench tectonics: sediment thickness across Philippine Plate.

RESULTS AND DISCUSSIONS

The conducted statistical data modelling and several types of the regression analysis were aimed to compare the variance in the geological data sets of the Mariana Trench explained by the complex interplay of the geomorphic, geological and oceanological attributes of the data and the bathymetric factors of the location of several segments of the trench in various parts of the Pacific Ocean. Using StatsModels library of Python, in particular several linear models of the correlation between various factors were computed, analysed and explained by the groups of variables.

Tested environmental variables of the Mariana Trench include four main geological factors (location on the tectonic plates, slope steepness degree, sediment thickness of the layer and bathymetric depth and submarine volcanism) and attributes of the 25 cross-section bathymetric profiles (mean values, maximal depths, median values and two quartile sub-division of the data sets) and the shared variances between environment and attributes. Shared variance arises due to correlations between the factor of sediment thickness and slope angle degree, e.g. because the attributes of the sediment accumulation are influenced by the canyon shape apart from the directly or indirectly depending on the oceanological conditions of the submarine currents.

The graphical output shows normal data distribution as demonstrated by (Figure 2), where each QQ profile has a given value for data distribution by quantiles across the trench profiles. The distribution of the geologic residuals is shown on Figure 3 showing particularly data correlation for several cases: frequency of the data distribution by Pacific Plate, Philippine Plate, Mariana Plate, sediment thickness, and range of the bathymetric depths taken for the maximal values, and finally, geomorphological shape as slope angle degree. The results shown on Table 2 represent the computed numerical values of the previous graph (Figure 3).

The results on the Quantile regression (Figure 4) show the conditional median of the response geologic variable given changing bathymetric values with movement southwards across the Mariana Trench. Thus, the upper plots shows (Figure 4, A, B, C) data distribution across tectonic plates: Pacific Plate, Philippine Plate, Mariana Plate and Caroline Plate. The lower plot (Figure 4, D, E, F) shows data distribution for the Caroline Plate and cumulative sediment thickness and slope angle degree by profiles.

The numerical explanation of the Figure 4 with corresponding sub-plots is presented in Table 3. The Figure 5 shows dynamic regression model using Python function embedded in StatsModel: State Autoregressive Moving Average. Finally, Table 4 shows the Python code that was used to perform the procedure of SARIMAX and the resulting output table. The model shows autocorrelation of the data by bathymetric profiles. The results demonstrated a correlation between the geological variables and geospatial location of the samplings across Mariana Trench, which proves the interplay between multiple factors affecting its geomorphology.

CONCLUSIONS

While the usage of the traditional methods of geoinformatics and spatial analysis is, beyond doubts, strongly recommended for any research in geoscience, there is another powerful tool for the geospatial data processing other than GIS, sometimes overlooked or skipped by the geographers: a data modelling by use of Python or R programming languages. Python, an open source free programming language is highly suitable for the statistical analysis in geoscience research, since it has a powerful statistical and math libraries, e.g. StatsModel, highly effective for scientific computing and used in the current research. The functionality of Python language and StatsModel, tested in this work, is proved to be highly effective for the statistical analysis of the geo-marine sets.

The proposed approach of the Python based statistical analysis enables accurate and efficient computation and modelling of the large data sets in marine geology and oceanology. A challenge in the evaluation of geological big data sets (that is, several thousand of observation points as in this case, 12.590 samples) concerns the difficulty in manual identifying a correct algorithms in the computations and data distribution analysis.

The necessity to apply a precise machine learning algorithms is recently increased in geographic sciences with respect to the importance of choosing an effective method for data visualization and computation. The solutions to the mentioned above problems are provided by Pandas data frames: for example, optimizing structure of the data, selecting the correct parts from the whole data frame (columns, rows in the data arrays) and plotting. Based on the presented results, the application of Python programming language is strongly recommended in geoscience research as an addition to the traditional GIS methods.

Conflict of Interests: The author declares no conflict of interest.

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

REFERENCES

Ansley, C. F. & R. J. Kohn. (1985). Estimation, filtering, and smoothing in state space models with incompletely specified initial conditions. *Annals of Statistics* 13, 1286–1316. [CrossRef]

Bello-González, J. P., Contreras-Reyes, E. & Arriagada, C. (2018). Predicted path for hotspot tracks off South America since Paleocene times: Tectonic implications of ridge-trench collision along the Andean margin. *Gondwana Research*, 64, 216–234. [CrossRef]

Boston, B., Moore, G. F., Nakamura, Y. & Kodaira, S. (2017). Forearc slope deformation above the Japan Trench megathrust: Implications for subduction erosion. *Earth and Planetary Science Letters*, 462, 26–34. [CrossRef]

Björck, A. (1996). Numerical methods for least squares problems. SIAM, Philadelphia. ISBN 0-89871-360-9. [CrossRef]

Box, G. E. P.; Tiao, G. C. (1992). *Bayesian Inference in Statistical Analysis.* New York: John Wiley and Sons. ISBN 0-471-57428-7. (Section 8.1.1). [CrossRef]

Dierssen, H. M. & Theberge Jr. A. E. (2014). Bathymetry: History of Seafloor Mapping. Encyclopedia of Natural Resources, Taylor & Francis. [CrossRef]

Fujie, G., Ito, A., Kodaira, S., Takahashi, N., & Kaneda, Y. (2006). Confirming sharp bending of the Pacific plate in the northern Japan trench subduction zone by applying a traveltime mapping method. *Physics of the Earth and Planetary Interiors*, 157, 72–85. [CrossRef]

Grand, S. P., Hilst, R. D. van der, & Widiyantoro, S. (1997). Global Seismic Tomography: A Snapshot of Convection in the Earth. *GSA Today*, 7(4), 2–7.

Lemenkova, P. (2019). Processing Oceanographic Data by Python Libraries NumPy, SciPy And Pandas. *Aquatic Research*, 2(2), 73-91. [CrossRef]

Ljung, G. M. & Box, G. E. P. (1978). On a Measure of a Lack of Fit in Time Series Models. *Biometrika*. 65 (2): 297–303. [CrossRef]

Michibayashi, K., Tasaka, M., Ohara, Y., Ishii, T., Okamoto, A., & Fryer, P. (2007). Variable microstructure of peridotite samples from the southern Mariana Trench: Evidence of a complex tectonic evolution. *Tectonophysics*, 444, 111–118. [CrossRef]

Millman, K. J. & Aivazis, M. (2011). Python for Scientists and Engineers, *Computing in Science & Engineering*, 13, 9-12. [CrossRef]

Oliphant, T. (2015). *Guide to NumPy* (2 ed.). CreateSpace. ISBN 978-1517300074.

Oliphant, T. E. (2007). *Python for scientific computing*. Computing in Science & Engineering 9(3), 10-20. [CrossRef]

Reid, W. D. K., Cuomo, N. J., & Jamieson, A. J. (2018). Geographic and bathymetric comparisons of trace metal concentrations (Cd, Cu, Fe, Mn, and Zn) in deep-sea lysianassoid amphipods from abyssal and hadal depths across the Pacific Ocean. *Deep-Sea Research Part I*, 138, 11–21. [CrossRef]

Schellart, W. P. (2008). Subduction zone trench migration: Slab driven or overriding-plate-driven? *Physics of the Earth and Planetary Interiors*, 170, 73–88. [CrossRef]

Seabold, S. & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. Proceedings of the 9th Python in Science Conference.

Smith, W. H. F., & Sandwell, D. T. (1997). Global Sea Floor Topography from Satellite Altimetry and Ship Depth Soundings. *Science*, 277, 1956–1962. [CrossRef]

Strutz, T. (2016). *Data Fitting and Uncertainty (A practical introduction to weighted least squares and beyond).* Springer Vieweg. ISBN 978-3-658-11455-8.

Taira, K., Yanagimoto, D., & Kitagawa, S. (2005). Deep CTD Casts in the Challenger Deep, Mariana Trench. *Journal of Oceanography*, 61, 447–454. [CrossRef]

Theberge, A. (2008). Thirty years of discovering the Mariana Trench. *Hydro International*, 12, 38–39.

Timm, N. H. (2007). *Applied Multivariate Analysis.* Springer Science & Business Media, 695 p. ISBN: 978-0-387-95347-2. [CrossRef]