doi.org/10.54473/IJTRET.2024.8106

International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 1 February 2024
ISSN No.2582-0958

MAPPING AND PREDICTION OF LAND SUBSIDENCE

[1]Aaradhyanidhi Aiyer, [1] Reshma Refina Khan, [1]Abinayasri M, [2]Mr Jeya ganesan J
[1] UG Students, [2]Asst.Professor, Dept. of Artificial Intelligence & Data Science
Sri Sairam Institute of Technology, Sai Leo Nagar, West Tambaram, Chennai-44

Received 02 March 2024 Received in revised form 06 March 2024 Accepted 10 March 2024

ABSTRACT

The exact location and timing of a land subsidence-related calamity cannot be foreseen with any degree of confidence. This is true for both progressive subsidence caused by fluid (groundwater, oil or gas) withdrawal and abrupt subsidence on surface caused by underground mine collapse. The best way to deal with these dangers is to mitigate them. In an ideal world, all regions prone to such risks would be well-known, and measures would be made to either prevent causing the problem if it is human-caused, or to avoid inhabiting such areas if they are prone to natural subsidence. Extensometers, levelling, hydrogeology modelling, and GPS are common methods for monitoring subsidence, although they require precise field data and are time-consuming. PSI (Persistent Scatterer Interferometry) is a strong radar-based remote sensing technique for measuring and monitoring surface displacements over time. The techniques of PSI using microwave sensors has made monitoring of earth deformation precisely more reliable and generates large number of Persistent Scatterer Candidates (PSC's) or sampling points. Although, all the above techniques have their own merits and de-merits but mapping and prediction of susceptible subsidence prone zones is an issue. This paper deals with the application of the Exploratory Data Analysis (EDA), a powerful environment in Data Science provided a set of tools that made it easier to interpret PSI processed Big Data and also improved the capability of validation, management, visualization, and presentation of results with greater reliability.

Keywords: Subsidence, Persistent Scattered Interferometry, Data Science, Exploratory data analysis, Visualization.

1. INTRODUCTION

The exact location and timing of a land subsidence-related calamity cannot be foreseen with any degree of confidence. This is true for both progressive subsidence caused by fluid (groundwater, oil or gas) withdrawal and abrupt subsidence on surface caused by underground mine collapse. The best way to deal with these dangers is to mitigate them. In an ideal world, all regions prone to such risks would be well-known, and measures would be made to either prevent causing the problem if it is human-caused, or to avoid inhabiting such areas if they are prone to natural subsidence. Extensometers, levelling, hydrogeology modelling, and GPS are common methods for monitoring subsidence, although they require precise field data and are time-consuming. PSI (Persistent Scatterer Interferometry) is a strong radar-based remote sensing technique for measuring and monitoring surface displacements over time. The techniques of PSI using microwave sensors has made monitoring of earth deformation precisely more reliable and generates large number of Persistent Scatterer Candidates (PSC's) or sampling points. Although, all the above techniques have their own merits and de-merits but mapping and prediction of susceptible subsidence prone zones is an issue. Although, all the above techniques have their own merits and de-merits but mapping and prediction of susceptible subsidence prone zones is an issue. This paper deals with the application of the Exploratory Data Analysis (EDA), a powerful environment in Data Science provided a set of tools that made it easier to interpret PSI processed Big Data and also improved the capability of validation, management, visualization, and presentation of results with greater reliability.

Geosciences and remote sensing are heavily reliant on data analysis at the moment [1]. In both academic and professional settings, the fields of data science and analytics are rapidly ascending in importance [2]. Data science is the study of cleaning, preparing, and analysing unstructured, semi-structured, and structured data, as well as combining it with other data [3]. Statistical, mathematical, programming, problem-solving and data capturing are all used in this scientific sector to get the most out of the data. EDA is a step in Data Science that examines your cleansed data to see if statistical processing is appropriate for a particular study. Big Data challenges affect a wide

range of industries and sectors, from economics and commerce to public administration, national security, and scientific research, among others [4]. R and Python are well-known data mining and analysis programming languages. High levels of dynamic and interactive nature, as well as a wealth of scientific libraries, make these languages suitable for analytical jobs. One of Python's most useful characteristics is its flexibility when it comes to visualizing and plotting data. Python's graphics utilities make it simple to present displays, surfaces, volumes, vector fields, histograms, animations, and a wide range of other data plots. The Python ecosystem has grown as a result of this. Python's data visualization libraries include Pygal, Altair, VisPy, PyQtGraph, Matplotlib, Bokeh, Seaborn, Plotly, and ggplot, which are all commonly used for charting data[3]. For making colorful statistical charts, Python Seaborn is excellent. This is especially true for web-based interactive presentations, where Python Bokeh shines.

After extensive literature review[2,4-6] the main objective of this study is to apply a novel approach of data science techniques on the Persistent Scatterer Interferometric (PSI) processed Big Data using exploratory data analysis for mapping and prediction of land subsidence.

2. MATERIALS AND METHODS

The pre-processed PSI data in .csv format using 100 images of Sentinel-1A satellite from 2015-2018 with refined 685 subsidence Persistent Scatterer Candidates (PSC's) or sampling points was acquired for the study area of 23 km^2 lying within geographical coordinates of 86.45° longitude and 22.53° latitude in the state of Jharkhand, India. The reasons of land subsidence is beyond the scope of this paper and only deals with visualization of results using Data Science techniques. The application of Data Science visualization techniques was carried out on Python script in Anaconda software distribution using the Juypter notebook, an open-source, web-based IDE (Integrated Development Environment). Finally, the results obtained from EDA were digitized in an ArcGIS software and exported into .kml file for better interpretation and visualization of results in Google Earth platform. The complete flow chart of the methodological approach adopted to accomplish the objectives is given below (Figure 1).

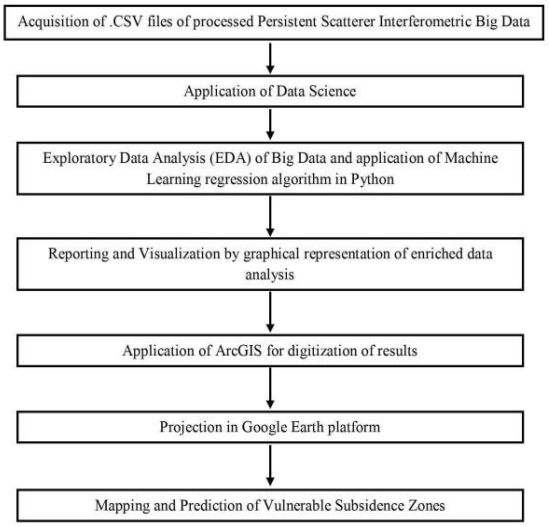

Figure 1: Flow chart of methodological approach

3. EXPLORATORY DATA ANALYSIS

Exploratory Data Analysis (EDA) is a type of data analysis in which you examine your cleaned data and apply statistical processing to it for specific analysis purposes. Python and R are well-known data mining and analysis programming languages. Because of their high levels, dynamic, and interactive nature, as well as the amount of scientific libraries, these languages are attractive for analytical professions. One of Python's most significant features is its ability to display and plot data rapidly and in a variety of ways. Displays, surfaces, volumes, vector fields, histograms, animations, and a variety of other data plots are all easily displayed with Python's graphics utilities. These have led to the development of the Python ecosystem of libraries. Data scientists use exploratory data analysis (EDA) to analyse and investigate data sets and summarise their main characteristics, often using data visualisation methods. It aids in determining how to best manipulate data sources to obtain the answers required, making it easier for data scientists to discover patterns, detect anomalies, test hypotheses, and validate assumptions. EDA is primarily used to discover what data can reveal beyond the formal modelling or hypothesis testing tasks, and it provides a better understanding of data set variables and their relationships. The python codes used for the analysis of the Big Data is given below:

```
#import pandas as pd; import numpy as np; import matplotlib.pyplot as plt
%matplotlib inline; import seaborn as sns
#dfCDSRF=pd.read_csv('E:\\EDA-Analysis\\CDSRF.csv')
```

```
In [6]: dfCDSRF.info()

<class 'pandas.core.frame.DataFrame'>
RangeIndex: 685 entries, 0 to 684
Data columns (total 7 columns):
PS-ID          685 non-null int64
LAT            685 non-null float64
LON            685 non-null float64
VEL            685 non-null float64
SIGMA VEL      685 non-null float64
CUMUL.DISP.    685 non-null float64
COHER          685 non-null float64
dtypes: float64(6), int64(1)
memory usage: 37.6 KB
```

The scatter plot is used to identify outlier PSC's with maximum velocity (Figure .2).

```
In [18]: plt.scatter(x='VEL',y='PS-ID',c='r', data=dfCDSRF)
         plt.xlabel('VEL (mm/yr)')
         plt.ylabel('PS-ID')
         plt.title('Graph in 2D')

Out[18]: Text(0.5, 1.0, 'Graph in 2D')
```

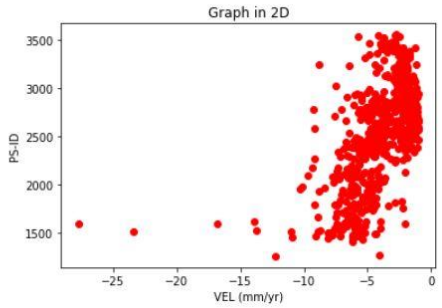

Figure 2: PS outliers with max. velocity

Application of the Machine Learning

Regression analysis is a fundamental concept in the field of Machine Learning (ML). The Machine Learning algorithm of Linear regression and Lowess regression was done using a seaborn scatter plot (Figure 3.2-3.3) to understand the trend of subsidence over the years. For determining the relationship between two continuous variables, simple linear regression is effective. The correlation coefficient is a measurement of the degree of linear relationship between two continuous variables, or how close the scatter of points is to a straight line when plotted together. A scatter plot is a helpful tool for examining the relationship between two continuous variables. The observed values of two variables are shown as points on a coordinate grid in a scatter plot. One of the variables' values aligns with the horizontal axis, while the other variable's values align with the vertical axis. The plotted points show a pattern that represents the link between these two variables.

There are a variety of statistical techniques that may be used to determine how one dataset is linked to another. Correlation analysis is the technical phrase for determining such a link. Although there are various methods for determining correlation, the Pearson r or simple linear correlation is the most often employed.

To determine the strength of a linear relationship between two continuous variables, the Pearson Correlation Coefficient is used. The Pearson's correlation coefficient, when applied to a sample, is commonly represented by r_{xy} and may be referred to as the sample correlation coefficient or the sample Pearson correlation coefficient. Given paired data $\{(x_1,y_1),\ldots\ldots(x_n,y_n)\}$ consisting of n pairs, r_{xy} is defined as:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}}$$

Where,
r_{xy} is the correlation coefficient
n is the sample size
x_i,y_i is the individual sample points indexed with i

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ is the sample mean of the x-variable;

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ is the sample mean of the y-variable

Therefore, can be written as:

$$r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2-(\sum x_i)^2}\sqrt{n\sum y_i^2-(\sum y_i)^2}}$$

or, $$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{\sqrt{\sum x_i^2 - n\bar{x}^2}\sqrt{\sum y_i^2 - n\bar{y}^2}}$$

An equivalent expression, as the mean of products of the standard scores as follows:

$$r_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right)$$

Where, $\left(\frac{x_i-\bar{x}}{s_x}\right)$ is the standard score of x and similar for y.
Alternatively,

International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 1 February 2024
ISSN No.2582-0958

$$r_{xy} = \frac{\sum x_i y_i - n\bar{x}\bar{y}}{(n-1)\, s_x\, s_y}$$

Where,

$s_x = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$ is the sample standard deviation and similar for S_y.

The python codes for the regression analysis is given below:

```
In [25]: plt.rcParams.update({'figure.figsize':(10,8), 'figure.dpi':100})
         sns.lmplot(x='LON', y='LAT', data=dfCDSRF)
         plt.title("Scatter Plot with Linear fit")
Out[25]: Text(0.5, 1, 'Scatter Plot with Linear fit')
```

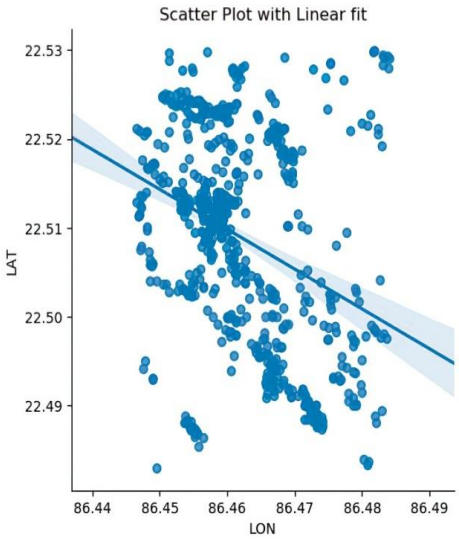

Figure 3: Linear regression analysis of data

```
In [20]: plt.rcParams.update({'figure.figsize':(10,8), 'figure.dpi':100})
         sns.lmplot(x='LON', y='LAT', data=dfCDSRF, lowess=True)
         plt.title("Scatter Plot with Lowess fit")
Out[20]: Text(0.5, 1, 'Scatter Plot with Lowess fit')
```

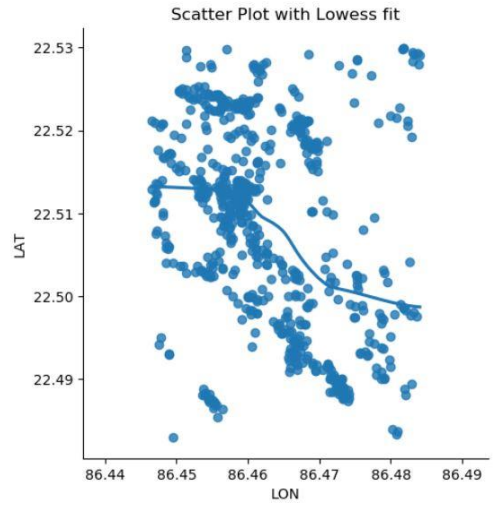

Figure 4: Lowess regression analysis of data

Simple Linear Regression is a statistical method that allows us to summarize and study relationships between variables. In case of linear regression, a straight line relationship is fitted even between two continuous variables. Since latitude and longitude can take any value in the available range its clear that they are continuous variables. LOWESS (Locally Weighted Scatter Plot Smoothing), sometimes known as LOESS (locally weighted smoothing), is a common regression analysis method that draws a smooth line through a time plot or scatters plot to reveal relationships and predict trends (Figure 4).

The Machine Learning (ML) algorithm of regression analysis with the recent 2015-2018 Sentinel-1A data confirms the continuing deformation trend in the study area. The regression analysis defines the pattern of the density of subsidence PS's during the observation period. It also exposes the subsidence vulnerability of the region around the regression line in geographical coordinates. The histogram analysis of the subsidence PSC's with Temporal Coherence (Figure 5) Cumulative displacement (mm) (Figure 6) and Displacement Velocity (mm/yr) (Figure 7) demonstrates the reliability and accuracy of the results. The displacement velocity histogram (Figure 7) also demonstrates the PSC outliers and cluster of PSC's showing slow subsidence velocities.

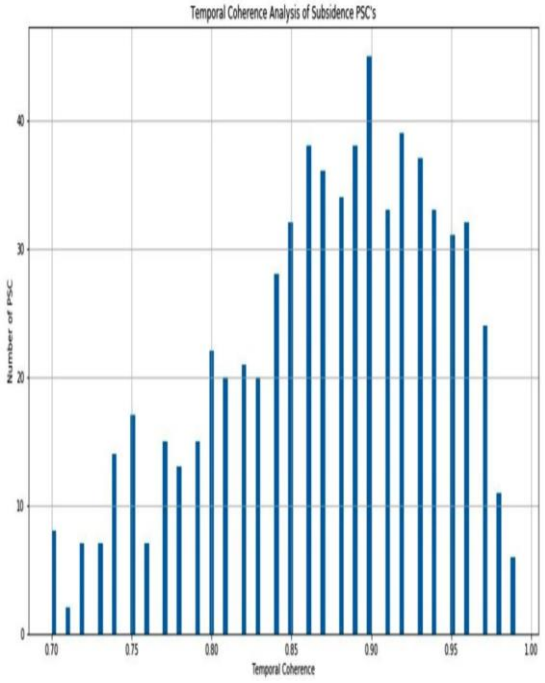

Figure 5: Distribution of Temporal Coherence of PS's

International Journal of Trendy Research in Engineering and Technology
Volume 8 Issue 1 February 2024
ISSN No.2582-0958

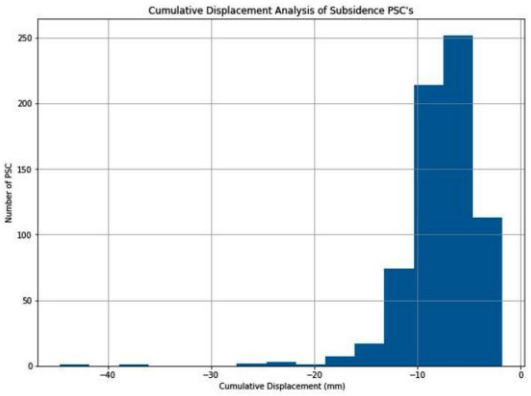

Figure 6: Analysis of Cumulative Displacement (mm)

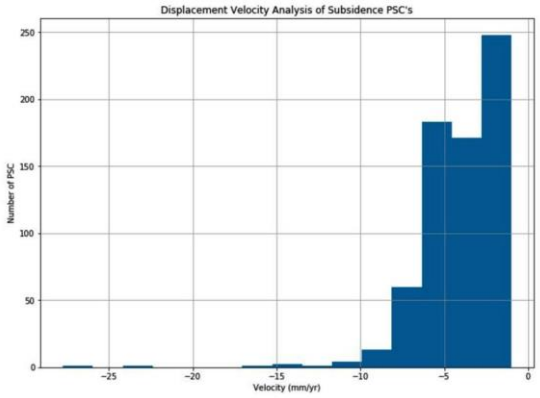

Figure 7: Analysis of Displacement Velocity (mm/yr)

Kernel Density Estimation (KDE) is a non-parametric method of estimating a random variable's Probability Density Function (PDF) in statistics. Gaussian kernels are used in this function, which also contains automatic bandwidth determination.

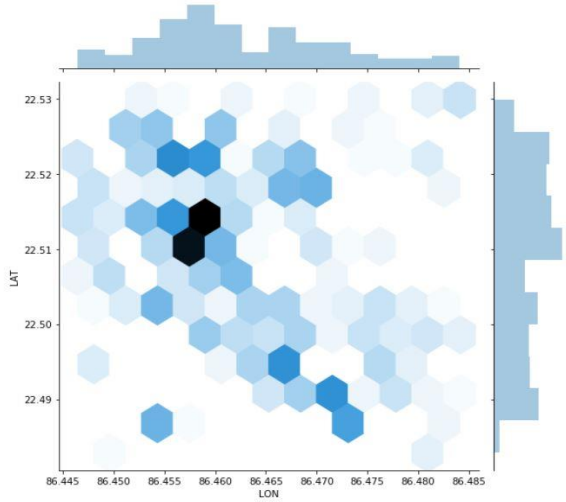

Figure 8:Active deformation zones (the darker hexagon indicates the density of subsidence PS's)

The Seaborn joint plot is used to generate a 2D PS kernel density map to locate active deformation zones, the darker hexagon indicates the density of negative velocities in the area and finally, the zone demarcation map (Figure 8) and active zones demarcated based on the density of subsidence PSC's (Figure 9) on geographical coordinates were generated to interpret the most susceptible land deformation zones.

The subsidence vulnerability of the study area is demarcated and mapped based on the density of subsidence PS's. As observed, the subsidence phenomenon is quite prominent and is widely spread throughout the study are

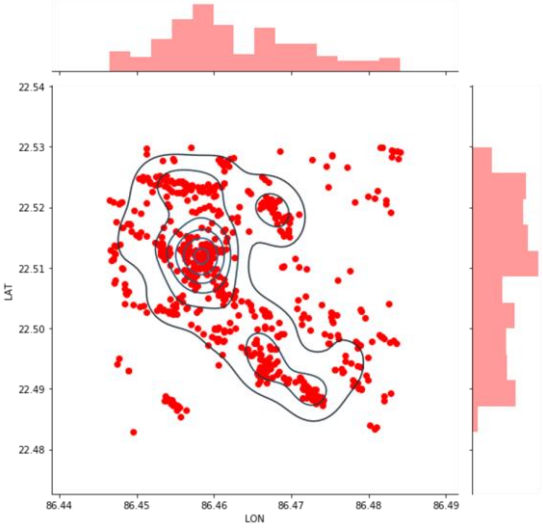

Figure 9: Active zones demarcated based on the density of subsidence PSs

Mapping and Prediction of Vulnerable Land Subsidence Zones

The ArcGIS platform is used to georeference and digitize the results obtained from the application of the EDA technique.

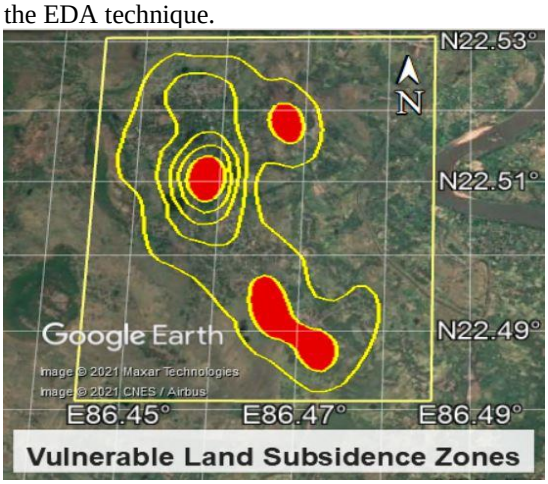

Figure 10: Mapping of Predicted Vulnerable

Land Subsidence Zones

The digitized results were projected on Google Earth platform. The final mapping of predicted vulnerable land subsidence zones is done and shown below (Figure 10).

4. RESULTS AND DISCUSSION

The ML algorithm of regression analysis defines the pattern of the density of subsidence PS's during the observation period. It also exposes the subsidence vulnerability of the region around the regression line in geographical coordinates. Slow deformation is evident with negative velocities between −27.74 to −1 mm/yr. The cumulative displacement varies from -44.64 mm to -1.79 mm during the observation period. However, it is quite evident that the majority of PSC's are forming a cluster, showing the present deformation trend, with a velocity between -2 to -10 mm/yr (Figure 3.1, 3.5 & 3.6), indicating a slow and continuous deformation over the region. Few PS outliers are showing higher negative deformation velocity in the study area, particularly PSID-1595 with -27.74 mm/yr, PSID-1515 with -23.40 mm/yr and PSID-1597 with -16.78 mm/yr, which needs a thorough investigation.

5. CONCLUSIONS

Data Science visualization techniques using processed PSI data has immensely extracted the unknown facts embedded in the Big Data. Thereby, exposing new frontiers for research. Data Science visualization tools has substantially improved the interpretation and visualization of the results, which has opened up new study avenues. Land deformation could be monitored in real-time utilising more modern satellite sensors in the Big Data age and Big Data analytics. The mapping of predicted vulnerable land subsidence zones, finally confirms that the study site is suffering from subsidence movements, which is likely to affect the life of inhabitants in the future. The future of land deformation monitoring and prediction appears to be in the hands of Artificial Intelligence and Machine Learning/Deep Learning.

ACKNOWLEDGEMENTS

The author is highly obliged and thankful to the HOD, Dept. of Artificial Intelligence & Data Science, Sri Sairam Institute of Technology, for the encouragement and support.

Conflict of Interest

The author declare no conflict of interest.

REFERENCES

[1]. Lary, D. J., Alavi, A. H., Gandomi, A. H., & Walker, A. L. (2016). Machine learning in geosciences and remote sensing. *Geoscience Frontiers*,7(2016)3-10 doi.org/10.1016/j.gsf.2015.07.003

[2]. Rogel-Salazar, J. (2018). Data science and analytics with python. In *Data Science and Analytics with Python*. CRC Press. doi.org/10.1201/9781315151670

[3]. Embarak, D. O. (2018). *Data Analysis and Visualization Using Python*. Apress. doi.org/10.1007/978-1-4842-4109-7

[4]. Philip Chen, C. L., & Zhang, C.-Y. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Information Sciences*, *275*(2014) 314–347.doi.org/10.1016/j.ins.2014.01.015

[5]. El Kamali, M., Abuelgasim, A., Papoutsis, I., Loupasakis, C., & Kontoes, C. A reasoned bibliography on SAR interferometry applications and outlook on big interferometric data processing. In *Remote Sensing Applications: Society and Environment*,19(2020) 100358 doi.org/10.1016/j.rsase.2020.100358

[6]. Minh, D. H. T., Hanssen, R., & Rocca, F. Radar interferometry: 20 years of development in time series techniques and future perspectives. *Remote Sensing*, *12*(2020) 1364. doi.org/10.3390/RS12091364