ახალგაზრდა მეცნიერთა საერთაშორისო კონფერენცია „დედამიწის შემსწავლელი მეცნიერების თანამედროვე პრობლემები“
შრომები, ISBN 978-9941-36-044-2, თბილისი, საქართველო, 21-22 ნოემბერი, 2022 წ.
International Conference of Young Scientists "Modern Problems of Earth Sciences"
Proceedings, ISBN 978-9941-36-044-2, Tbilisi, Georgia, November 21-22, 2022

EVALUATION OF NATURAL DISASTERS USING MACHINE LEARNING AND BIG DATA FOR GEORGIA

Palavandishvili A.

Institute of Hydrometeorology of Georgian Technical University, Tbilisi, Georgia

Abstract. *The purpose of the presented paper is to estimate and analyze the distribution of precipitation over the territory of Georgia to study drought related processes and other natural hazards, as well as to evaluate the possibility of environment hazard detection at the early stage of their evolution, monitoring and prediction. We intend in our study to apply the Machine Learning, the branch of computer science which focuses on the use of big data and algorithms to imitate the way that humans learn, gradually improving its accuracy. Such analysis allows developing a Combined Drought Index (CDI) and corresponding drought hazard 5km resolution map.*
Key words: *Machine Learning, Big data, natural disaster, drought*

Introduction

Economic and other losses from natural disasters are increasing throughout the world. According to the International Disaster Database (EM-DAT), over the last 70 years, hydro-meteorological disasters have shown the fastest rate of increase of all disaster types. In parallel, technological capabilities to manage such disasters have advanced rapidly.

Mitigating the impacts of climate change and successful adaptation requires effective climate change strategic planning by countries worldwide whose decision-making requires complex models and sources of information. The Big Data toolkit enables the systematization, processing, and evaluation of heterogeneous data and information sources, which is unfeasible with traditional disciplinary analysis tools. In Europe the Global Monitoring for Environment and Security (GMES) initiative of the European Commission and the European Space Agency (ESA) is actively supporting the use of satellite technology in disaster management, with projects such as PREVIEW (Prevention, Information and Early Warning pre-operational services to support the management of risks), LIMES (Land and Sea Integrated Monitoring for Environment and Security), GMOSS (Global Monitoring for Security and Stability), SAFER (Services and Applications For Emergency Response), and GMOSAIC (GMES services for Management of Operations, Situation Awareness and Intelligence for regional Crises) [1]. The United Nations Platform for Space-based Information for Disaster Management and Emergency Response (UN-SPIDER, 2010) has been established by the UN to ensure that all countries have access to and develop the capacity to use space- based information to support the disaster management cycle. They are working on a space application matrix that will provide the satellite-based approaches for each type of hazard and each phase of the disaster management cycle.

Data and Method

On order to carry out research observation net data of National Environmental Agency and Copernicus ERA5 precipitation reanalysis data are used. ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables. The data cover the Earth on a 30km grid and resolve the atmosphere using 137 levels from the surface up to a height of 80km. ERA5 includes information about uncertainties for all variables at reduced spatial and temporal resolutions.

Quality-assured monthly updates of ERA5 (1959 to present) are published within 3 months of real time. Preliminary daily updates of the dataset are available to users within 5 days of real time. ERA5 is the latest climate reanalysis produced by ECMWF, providing hourly data on many atmospheric, land-surface and sea-state parameters together with estimates of uncertainty. ERA5 data are available in the Climate Data

Store on regular latitude-longitude grids at 0.25o x 0.25o resolution, with atmospheric parameters on 37 pressure levels.

Climate Hazards Group InfraRed Precipitation with Station data (CHIRPS) is a 35+ year quasi-global rainfall data set. Spanning 50°S-50°N (and all longitudes) and ranging from 1981 to near-present, CHIRPS incorporates our in-house climatology, CHPclim, 0.05° resolution satellite imagery, and in-situ station data to create gridded rainfall time series for trend analysis and seasonal drought monitoring. The Integrated Multi-satellite Retrievals for GPM (IMERG) algorithm combines information from the GPM satellite constellation (https://gpm.nasa.gov/missions/GPM/constellation) to estimate precipitation over the majority of the Earth's surface. This algorithm is particularly valuable over the majority of the Earth's surface that lacks precipitation-measuring instruments on the ground**.** Standard Deviation is a measure which shows how much variation (such as spread, dispersion, spread,) from the mean exists. The standard deviation indicates a "typical" deviation from the mean. It is a popular measure of variability because it returns to the original units of measure of the data set.

Data visualization allows us to see availability and missing variables, for this purpose was used R-instat free software, which is easy to use and gives all necessary functions to analyze small amount of data. In order to compare two datasets, we used statistical method, in this case is pearsons correlation, mean absolute error and standard deviation.

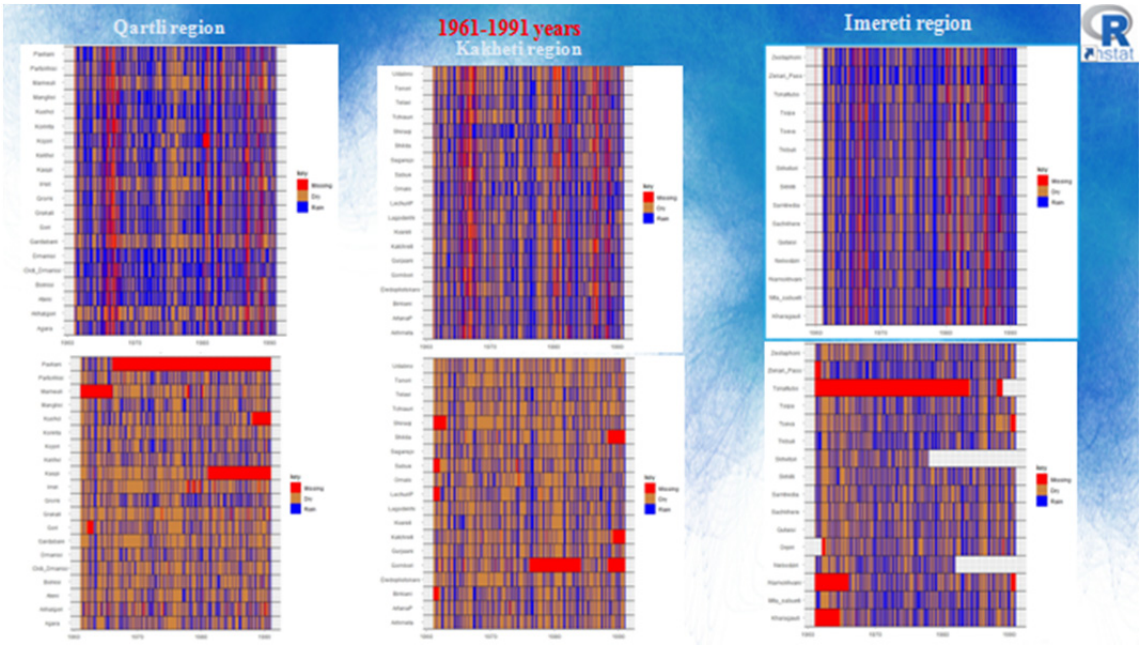

Fig.1. Inventory of station data of Qartli, Imereti anf kakheti regions and ERA5 data.

Machine learning is an important component of the growing field of data science. Through the use of statistical methods, algorithms are trained to make classifications or predictions, and to uncover key insights in data mining projects. These insights subsequently drive decision making within applications and businesses, ideally impacting key growth metrics. As big data continues to expand and grow, the market demand for data scientists will increase. They will be required to help identify the most relevant business questions and the data to answer them.

Overfitting is a concept in data science, which occurs when a statistical model fits exactly against its training data. When this happens, the algorithm unfortunately cannot perform accurately against unseen data, defeating its purpose. Generalization of a model to new data is ultimately what allows us to use machine learning algorithms every day to make predictions and classify data.

CHIRPS and IMG satellites data have been compared to station ones. The statistical analysis showed that for Caucasus region CHIRPS data are more valid.

Table 1. Statistical parameters of CHIRPS and station data for 2000-2020 year.

Statistical Method of Environment Data Analysis

Station	correlation	Mean absolute error	Standard deviation
Akhalqalaqi	0.66759	28.7	26.84318
Alpana	0.587625	41.8	46.65444
Ambrolauri	0.64799	27.6	35.83004
Goderdzi Pass	0.491516	46.3	45.11875
Gori	0.600696	18.8	23.40047
Gurjaani	0.68528	29.1	38.4494
Khaishi	-	-	-
Kharagauli	-	-	-
Khashuri	0.602859	20.7	25.35372
Khulo	0.668731	49.2	58.73967
Lagodekhi	0.597565	56.8	60.23253
Lentekhi	0.553973	44.3	53.99864
Luji	0.590198	42.4	54.02934
Mukhrani	0.670568	22.5	24.8989
Omalo	-	-	-
Oni	0.520496	36.2	39.74461
Shovi	0.717298	30.0	36.37001
Stephantsminda	-	-	-
Tbilisi	0.678032	21.1	28.35372
Telavi	0.693299	26.0	35.45118
Tetri-Tskaro	-	-	-
Tseva	0.607114	39.7	41.05641
Zestaphoni	0.462851	41.3	54.92459
Zugdidi	0.603897	51.9	62.13524

Standard deviation

$$\sigma = \sqrt{M[(X-\mu)^2]}$$

Pearson's correlation

$$\rho_{X,Y} \equiv Corr(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y} = \frac{M[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Discussion

At the first stage of data processing, it is necessary to remove those rows which have incomplete observation or the data does not correspond to reality. R-instat, recommended by the World Meteorological Organization (WMO), is used to visualize the data series.

R-instat is also used for data comparison, due to the inhomogeneity of the precipitation, three parameters are analyzed: correlation, standard deviation and Mean absolute error. For this study, a region particularly vulnerable to drought- Kakheti was selected, where the statistical parameters between the station and ERA are as follows: the best correlation was observed at Omalo station 0.65, and the lowest at Shiraki station - 0.30, the lowest standard deviation at Udabno station - 4.04 , the highest- 7.62 (MAE) -minimum at station Udabno-1.66, maximum at station Shilda -3.04.

The three month SPI-3 is calculated for data validation for both station and satellite values (https://edo.jrc.ec.europa.eu/) The SPI (standardized precipitation index) classifies the precipitation sums on a particular date with respect to the sums of the same month in all years of the measurement record. For this purpose, the precipitation sums of the whole record within one month around the respective date are transformed into a standard normal distribution around zero. The SPI is these transformed precipitation sums [3]. The SPI value directly indicates the frequency of the observed precipitation amount in the corresponding month as estimated from the whole observation record. Correlation analysis of these two data was conducted, the results obtained are lower than the original correlation, the reasons for this may be the following: satellite error, (the satellite perceives precipitation also solid precipitation), data break at the station, in this case the minimum correlation value fells down to -0.08 and the maximum increases up to 0.75 unit, of course, this index was recalculated for other periods, one month, too and the correlation value did not change, also another R-studio software was used to make sure the result reliability. In this case the correlation values did not change as well (CHIRPS satellite data were used, the 8-th month of 2007 year data are missing for all stations; There are no 2010 and 2011data at all).

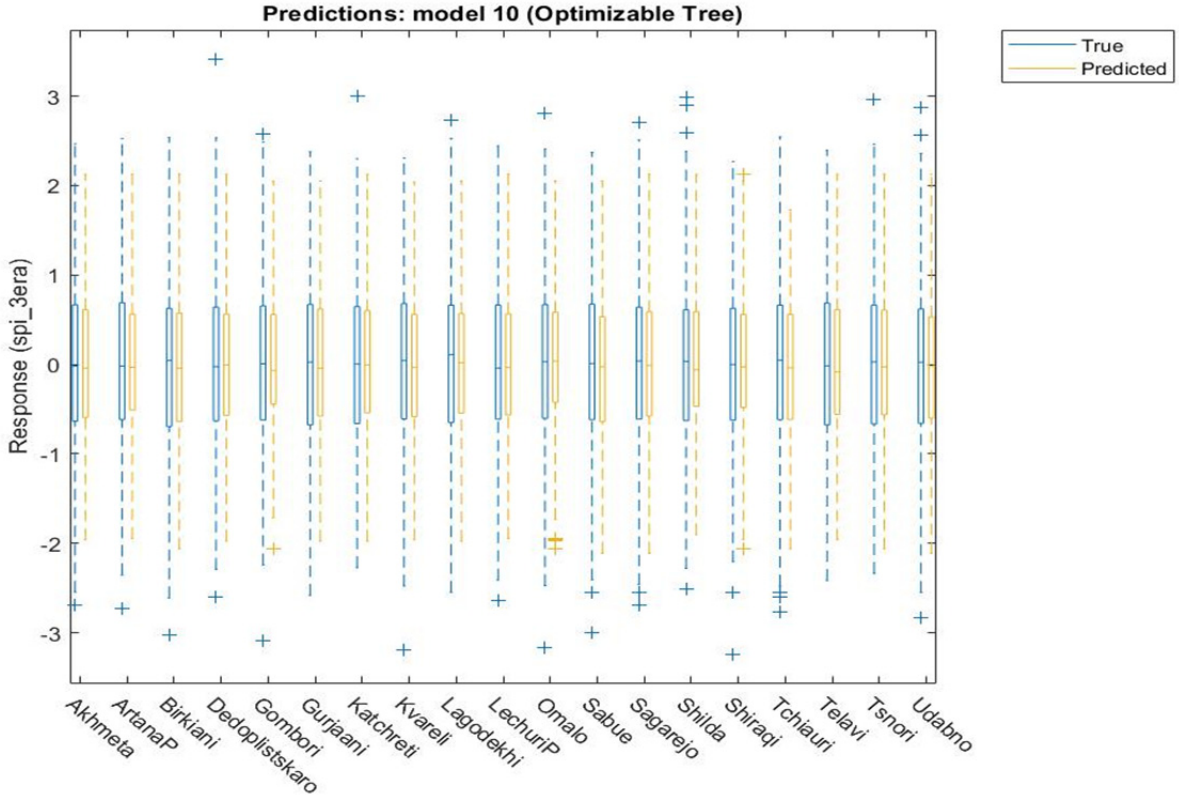

Fig 2. r-squared shows how well the SPEI3 data fit the regression model.

Considering that satellite observations have been produced, in the case of CHIRPS since 1987, and that station data has discontinuities and inaccuracies that are unfavorable for environmental hazard assessment, the need for machine learning has emerged. The first region that was subjected to analysis is Kakheti - 19 stations 1961-1990 30-year period data were taken. Based on these data, the Standardized Precipitation Index (SPI) for 3 months was calculated and then subjected to machine learning. In this case, the most optimal result was shown by the "optimized tree", where the minimum number of leaves is equal to 45, and the training time is 30,296 s, the prediction speed_~ 43000 obs/sec.

Despite such good parameters, it was necessary to add additional stations, because there was not enough information in the Kakheti region for the correct analysis of machine learning, comparing the model and station data, it was found that overfit was obtained, which is not valid for the analysis of new data, it was necessary to add the Kartli region as well, in this case it was also highlighted , that stations that are close to each other give better results to the model, while, in the case of both regions, stations that are far away increase the probability of overfit. Therefore, machine learning needs to increase the observation network, the number of stations, the observation period, as well as the selection of satellite data for the region where there is not enough observation.

Conclusion

A support vector machine (SVM) was selected in the Matlab space - an algorithm for supervised machine learning, which allows us to optimize it, for example, we can control the number of divisions in the "machine tree", which will help us achieve the accuracy of the model, as presented, the tenth model showed the best result, with the help of months At each point, we can determine the probability of drought.

It was necessary to add additional stations, because not enough information was found in the Kakheti region for the correct analysis of machine learning, comparing the model and station data it was found that overfit was obtained, which is not valid for the analysis of new data, it was necessary to add the Kartli

region, in this case it was also revealed that which are close to each other gave a better result for the model, while, in the case of both regions, stations that are further apart increase the probability of overfit. Therefore, machine learning needs to increase the observation network, the number of stations, the observation period, as well as select satellite data for the region where there is not enough observation. The following is necessary to avoid overfit:

- **Increase number of dataset**
- **Increase of time period**
- **Find nearest points of observations**

References

1. Tatishvili M., Palavandishvili, A. Tsitsagi M., Suknidze N. The Use of Structured Data for Drought Evaluation in Georgia//Journals of Georgian Geophysical Society, *25*(No.1) .DOI: https://doi.org/10.48614/ggs2520224806
2. Tsitsagi M., Tatishvili M., Gulashvili Z. CORRELATION OF DROUGHT INDICES FOR DIFFERENT CLIMATE CONDITION IN GEORGIA// Proceedings of INTERNATIONAL SCIENTIFIC CONFERENCE LANDSCAPE DIMENSIONS OF SUSTAINABLE DEVELOPMENT SCIENCE – CARTO/GIS – PLANNING – GOVERNANCE pp. 296-302
3. Palavandishvili A. Structured data set in environmental issues. The Regional Student Scientific Practical Conference (GTU-DAAD). 2022