
Quantitative Assessment of Drought Impacts Using XGBoost based on the Drought Impact Reporter

Beichen Zhang * **Fatima K. Abu Salem** † **Michael J. Hayes** * **Tsegaye Tadesse** *
bzhang25@unl.edu fa21@aub.edu.lb mhayes2@unl.edu ttadesse2@unl.edu

Abstract

Under climate change, the increasing frequency, intensity, and spatial extent of drought events lead to higher socio-economic costs. However, the relationships between the hydro-meteorological indicators and drought impacts are not identified well yet because of the complexity and data scarcity. In this paper, we proposed a framework based on the extreme gradient model (XGBoost) for Texas to predict multi-category drought impacts and connected a typical drought indicator, Standardized Precipitation Index (SPI), to the text-based impacts from the Drought Impact Reporter (DIR). The preliminary results of this study showed an outstanding performance of the well-trained models to assess drought impacts on agriculture, fire, society & public health, plants & wildlife, as well as relief, response & restrictions in Texas. It also provided a possibility to appraise drought impacts using hydro-meteorological indicators with the proposed framework in the United States, which could help drought risk management by giving additional information and improving the updating frequency of drought impacts. Our interpretation results using the Shapley additive explanation (SHAP) interpretability technique revealed that the rules guiding the predictions of XGBoost comply with domain expertise knowledge around the role that SPI indicators play around drought impacts.

1 Introduction

Drought is one of the most costly natural disasters in the world because of its broad impacts on various sectors in society [1]. Ongoing climate change is inclined to increase the frequency and intensity of drought by raising extreme variabilities over space and time in the hydrological cycle [2, 3, 4]. However, compared to other natural disasters, such as floods and wildfire, drought impacts often lack structural and visible existence. Thus, drought impacts on different socio-economic aspects could be either tangible or intangible, direct or indirect [3, 5]. Additionally, based on the propagation of a drought event, its impacts could last for weeks to even years [1]. The complex drought characteristics make it difficult to quantitatively monitor and evaluate drought impacts under climate change.

Many drought indicators have been developed to monitor drought intensity and frequency in recent decades. They are commonly grouped into meteorological, agriculture, hydrological, and composite drought indices based on the drought types [6, 7, 8]. However, only a few studies have tried to connect and calibrate drought indicators to various drought impacts, although we need to transform the temporal and spatial information of drought intensity and frequency into its impacts to help people and government agencies proactively prepare and mitigate drought [9].

Overall, two main challenges for quantitatively evaluating drought impacts are: 1. the complexity and non-linearity of relationships between drought indicators and drought impacts, and 2. scarcity of quantitative impacts data in multiple sectors with high quality and spatial and temporal resolution.

* School Of Natural Resources, University of Nebraska-Lincoln, Lincoln, NE, United States

† Computer Science Department, American University of Beirut, Beirut, Lebanon

To address those challenges, several studies employed text-based data sets, such as the European Drought Impact Report Inventory (EDII), as well as various regression models, in order to estimate the drought impacts [10, 11, 12, 13, 14, 15]. However, only one of the studies employed a machine-learning model (Random Forest) and results indicated that it had a better performance than regression models for describing complex drought impacts [13]. Additionally, because of the complexity and non-linearity of drought, the observations of its impacts are associated with the social vulnerability and resilience of the local ecosystem and environment [16]. Hence, an imbalanced sample distribution is common when the drought impacts data are collected, especially in the extreme categories, such as fire. Because of the potentially extreme cost of drought impacts, using a modeled approach to predict impacts will improve proactive drought response management. Overall, there is a need to develop a study with a systematic machine learning framework to link the imbalanced and multi-dimensional drought impacts with the drought indicators.

In this paper, we propose a machine-learning framework to predict multi-category drought impacts based on drought indices and impact reports in the United States, which to the best of our knowledge presents the first such attempt in the drought studies. The framework was developed based on the extreme gradient model (XGBoost) and tested by a case study in Texas during an identified drought period. We further interpret our machine learning algorithm using the Shapley additive explanation (SHAP), in order to render results that can be deemed trustworthy by domain experts.

2 Data and Methods

2.1 Data

We acquired the text-based drought impacts data set from the Drought Impact Reporter (DIR), developed and maintained by the National Drought Mitigation Center (NDMC). The DIR collects drought impacts from multiple resources and organizes them into nine categories¹. Monthly Standardized Precipitation Index (SPI) in the Pearson Type III distribution at the various temporal scales (1, 3, 6, 9, and 12 months) were generated based on a 30-year precipitation record. The precipitation data set was acquired from the Climate Hazards Group InfraRed Precipitation with Station (CHIRPS) [17]. Seasons and months were added to the predictors for counting the seasonality and temporal trend. Additionally, to describe the spatial characteristics of drought impacts, we employed the following geographic data sets in the case study: Land Cover (LC), Public Health Regions (PHR), Regional Water Project and Development (RWPD), and Texas A&M AgriLife Extension Service Districts (TAESD). All seasons, months, and spatial districts are categorical data sets.

2.2 The Proposed Framework

The framework of evaluating multi-dimensional drought impacts was developed based on the XG-Boost model, combining the drought monitoring with a typical machine-learning pipeline.

Data preparation and feature engineering. The drought impacts from the DIR were quantitatively summarized by month and converted to dummy variables (presence versus absence). The precipitation records were aggregated monthly and calculated to SPI1-12. We applied one-hot encoding on categorical data to remove the numerical categories and their effects on the model. All data were processed at the county level and divided into training, validation, and test data sets by stratifying based on the sample distribution of the impacts.

Addressing imbalanced data. For the drought impacts with a significant skew in the distribution that the proportion of the positive class is smaller than 20%, we applied the Synthetic Minority Oversampling Technique (SMOTE) and Random Undersampling on the training data sets in order to balance the class distribution and increase the model's response to the minority class in an attempt to improve learning. This step significantly improved the F2 score and recall for the models with an imbalanced sample distribution, such as fire. We also incorporated elements of cost-sensitive learning in the training of XGBoost.

Train and validate XGBoost models. XGBoost is an efficient implementation of the gradient boosting decision tree that employs the second-order Taylor series to proximate the cost function and

¹The nine categories are agriculture, energy, plants & wildlife, society & public health, water supply & quality, business & industry, fire, relief, response & restrictions, and tourism & recreation

Table 1: Summary of models performance on predicting drought impacts in the test data set, and the ratio of impacts is the number of impacts versus the total samples.

Category of Drought Impacts	Ratio of Impacts	Evaluation		
		Accuracy	Recall	F2 Score
Agriculture	0.69	0.86	0.93	0.92
Plants & Wildlife	0.29	0.79	0.79	0.74
Society & Public Health	0.50	0.90	0.96	0.94
Water Supply & Quality	0.36	0.78	0.51	0.55
Fire	0.11	0.88	0.80	0.68
Relief, Response & Restrictions	0.36	0.85	0.72	0.74

adds the regularization term in the objective function [18]. The reasons why we selected XGBoost to build models are five-fold. 1. Rule-based models such as decision-tree-based models are generally better suited than deep learning algorithms considering that our data set is of moderate size. 2. XGBoost models are convenient to build, in that they can attain highly-optimized performance by following standard hyperparameter search techniques implemented using stratified k-fold cross validation resampling during the training phase. XGBoost also can be easily trained in such a way to reduce overfitting. 3. XGBoost has been used successfully for winning several machine learning competitions. 4. Previous drought studies have used XGBoost for predicting meteorological indicators [19, 20]. 5. XGBoost can incorporate elements of cost-sensitive learning where a cost-matrix can help influence the model to produce less false negatives. We built and trained an XGBoost model for every selected category from the text-based drought impacts data. The binary cross-entropy loss function was applied in all of the models. Additionally, we tuned the following hyperparameters: gamma and the maximum depth of the XGBoost trees to activate pruning; lambda, the L2 regularization parameter; as well as the scale of positive weight that provides cost-sensitive training. A stratified 10-fold cross validation was used to validate the model’s stability on the validation data set after fine-tuning. The cross validation was designed in such a way so as to choose the hyperparameters that optimized the F2 score and area under the precision-recall curve (PR AUC). The latter two metrics emphasize that the recall values of our model are more important than precision.

Test and interpret models. To evaluate the models’ performance, we calculated the F2 score, recall, and accuracy on the test data set. The F2 score is the F-beta measure when beta is equal to 2 so as to increase the weight of recall. The F2 score and recall were selected because we considered the false negatives more costly than the false positives in predicting drought impacts. Additionally, the SHAP was employed to estimate each feature’s contribution to the model and explain the interactions among features.

Our work was written using Python 3.7 with the packages: scikit-learn, xgboost, numpy, pandas, gdal, netcdf4, xarray, geopandas, and climate_indices.

3 Results and Discussion

To examine the framework, we developed a case study for one of the most severe droughts in Texas from October 2010 to June 2015¹. Energy, business and industry, and tourism and recreation were dropped from the category because they accounted for less than 5% of the drought impacts during the period. Table 1 summarizes the proposed framework’s performance for assessing multi-dimensional drought impacts on the test data set. Except for water supply and quality (0.78) and plants and wildlife (0.79), the rest models’ accuracies ranged from 0.85 to 0.90. The water supply and quality model also had the lowest recall (0.51) and the F2 score (0.55). If we exclude this model, the rest of the models had recall values ranging from 0.72 to 0.96 and F2 score from 0.68 to 0.94. The fire model has the largest difference between the recall and the F2 score because we sacrificed the precision to gain a higher recall. Further studies are required to investigate why the water supply and quality model poorly performed. Overall, the framework has a good performance on evaluating various drought impacts using hydro-meteorological drought indices.

¹The period was identified based on the time-series plot from the United State Drought Monitor.

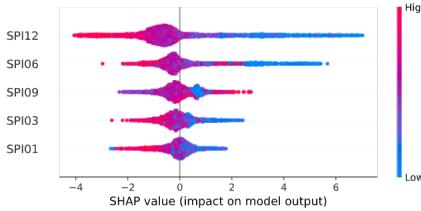


Figure 1: SHAP summary plot for SPI.

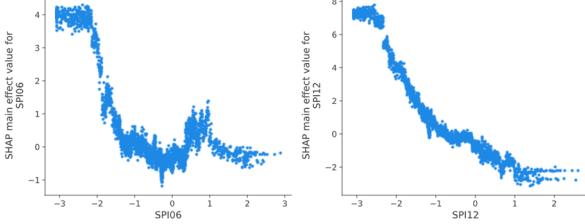


Figure 2: SHAP main effect plot for SPI6 and SPI12.

We now try to interpret the best-performing model using SHAP. Figure 1 shows the SHAP summary plot for drought indicators. Since the SHAP explainer has no explicit support at interpreting the one-hot encoded categorical data sets, we dropped them from the plot. The order of the features reveals their contributions to the model of society and public health impacts. The SPI with a 12-month moving window has the most significant impact on the model, followed by SPI6 and SPI9. SPI1 has the lowest impact on the model. Besides, most positive contributions result from the negative SPI values, except for some cases in SPI6 and SPI9, where some positive SPI values positively impact the model. This is in line with domain knowledge and expertise as follows. Typically, prolonged drought events are likely to lead to notable impacts on society and public health. Quick and minor drought events may not have any significant effects in this particular category because of the resilience in the human dimension. While SPI12 could enhance the signal of severe and prolonged droughts, one would expect it would have a more significant role in the model. However, future studies are required to explore why some higher positive SPI values would positively influence the drought impacts on society and public health. A possible inference is the quality of the drought impact data sets. Figure 2 is the SHAP main effect plot for SPI6 and SPI12 with the largest impacts on the model. The two scatter plots have a similar trend: the lowest values have the largest impacts on the model, while the impact on the model decreases with the SPI values increasing. It indicates that lower negative values in SPI6 and SPI12 would increase the probability of drought impacts on society and public health. However, further studies need to be done to explain the minor peak where SPI6 is around 0.8.

4 Conclusion and Future Work

Quantitatively identifying the various drought impacts is always a challenge to researchers. However, it is critical to transform and connect the temporal and spatial information from hydro-meteorological drought indicators to different drought impacts. This paper proposed an XGBoost-based framework to assess multi-category drought impacts with SPI and text-based data from the DIR. The framework has a good performance on the case study in Texas. The accuracy from the models running on test data sets ranged from 0.78 to 0.90, and the F2 score was from 0.55 to 0.94. We also explained and discussed the model for the drought impact on society and public health by applying the SHAP explainer, which provides a novel insight of drought impacts on society and public health. The results reveal that SPI12 had the greatest impacts on the society and public health model, and that negative SPI6 and SPI12 values might better explain the occurrence of drought impacts on society and public health more so than other indicators.

Further studies are recommended to investigate more profound reasons and linkages among the SPI indicators and drought impacts. Future work will focus on exploring the explanation of the categorical data sets and the model outputs from the other drought impacts. It is also worth applying the proposed framework to larger spatio-temporal data sets. This study opens the door to explore more machine-learning and deep-learning methods on converting information from the drought indicators about the intensity, frequency, and spatial extent to drought impacts.

References

[1] M. J. Hayes, M. D. Svoboda, B. D. Wardlow, M. C. Anderson, and F. Kogan, “Drought monitoring: Historical and current perspectives,” 2012.

[2] J. Quiggin, “Drought, climate change and food prices in australia,” *Melbourne: Australian Conservation Foundation*, 2010.

[3] Y. Ding, M. J. Hayes, and M. Widhalm, “Measuring economic impacts of drought: a review and discussion,” *Disaster Prevention and Management: An International Journal*, 2011.

[4] S. Mukherjee, A. Mishra, and K. E. Trenberth, “Climate change and drought: a perspective on drought indices,” *Current Climate Change Reports*, vol. 4, no. 2, pp. 145–163, 2018.

[5] A. F. Van Loon, T. Gleeson, J. Clark, A. I. Van Dijk, K. Stahl, J. Hannaford, G. Di Baldassarre, A. J. Teuling, L. M. Tallaksen, R. Uijlenhoet *et al.*, “Drought in the anthropocene,” *Nature Geoscience*, vol. 9, no. 2, p. 89, 2016.

[6] A. K. Mishra and V. P. Singh, “A review of drought concepts,” *Journal of hydrology*, vol. 391, no. 1-2, pp. 202–216, 2010.

[7] A. Zargar, R. Sadiq, B. Naser, and F. I. Khan, “A review of drought indices,” *Environmental Reviews*, vol. 19, no. NA, pp. 333–349, 2011.

[8] A. Dai, “Drought under global warming: a review,” *Wiley Interdisciplinary Reviews: Climate Change*, vol. 2, no. 1, pp. 45–65, 2011.

[9] S. Bachmair, C. Svensson, J. Hannaford, L. Barker, and K. Stahl, “A quantitative analysis to objectively appraise drought indicators and model drought impacts,” *Hydrology and Earth System Sciences*, vol. 20, no. 7, pp. 2589–2609, 2016.

[10] J. H. Stagge, I. Kohn, L. M. Tallaksen, and K. Stahl, “Modeling drought impact occurrence based on meteorological drought indices in europe,” *Journal of Hydrology*, vol. 530, pp. 37–50, 2015.

[11] S. Bachmair, I. Kohn, and K. Stahl, “Exploring the link between drought indicators and impacts,” *Natural Hazards and Earth System Sciences*, vol. 15, no. 6, pp. 1381–1397, 2015.

[12] V. Blauthut, L. Gudmundsson, and K. Stahl, “Towards pan-european drought risk maps: quantifying the link between drought indices and reported drought impacts,” *Environmental Research Letters*, vol. 10, no. 1, p. 014008, 2015.

[13] S. Bachmair, C. Svensson, I. Prosdocimi, J. Hannaford, and K. Stahl, “Developing drought impact functions for drought risk management,” *Natural Hazards and Earth System Sciences*, vol. 17, no. 11, pp. 1947–1960, 2017.

[14] M. M. de Brito, C. Kuhlicke, and A. Marx, “Near-real-time drought impact assessment: A text mining approach on the 2018/19 drought in germany,” *Environmental Research Letters*, 2020.

[15] T. Tadesse, B. D. Wardlow, J. F. Brown, M. D. Svoboda, M. J. Hayes, B. Fuchs, and D. Gutzmer, “Assessing the vegetation condition impacts of the 2011 drought across the us southern great plains using the vegetation drought response index (vegdri),” *Journal of Applied Meteorology and Climatology*, vol. 54, no. 1, pp. 153–169, 2015.

[16] D. A. Wilhite, M. D. Svoboda, and M. J. Hayes, “Understanding the complex impacts of drought: A key to enhancing drought mitigation and preparedness,” *Water resources management*, vol. 21, no. 5, pp. 763–774, 2007.

[17] C. Funk, P. Peterson, M. Landsfeld, D. Pedreros, J. Verdin, S. Shukla, G. Husak, J. Rowland, L. Harrison, A. Hoell *et al.*, “The climate hazards infrared precipitation with stations—a new environmental record for monitoring extremes,” *Scientific data*, vol. 2, no. 1, pp. 1–21, 2015.

[18] T. Chen and C. Guestrin, “Xgboost: A scalable tree boosting system,” in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[19] Y. Han, J. Wu, B. Zhai, Y. Pan, G. Huang, L. Wu, and W. Zeng, “Coupling a bat algorithm with xgboost to estimate reference evapotranspiration in the arid and semiarid regions of china,” *Advances in Meteorology*, vol. 2019, 2019.

[20] R. Zhang, Z.-Y. Chen, L.-J. Xu, and C.-Q. Ou, “Meteorological drought forecasting based on a statistical model with machine learning techniques in shaanxi province, china,” *Science of The Total Environment*, vol. 665, pp. 338–346, 2019.