GI_Forum 2022, Issue 1
Page: 58 - 76
Full Paper
Corresponding Author:
hannahkemper.speyer@gmail.com
DOI: 10.1553/giscience2022_01_s58

Development of a Drought Early Warning System based on the Prediction of Agricultural Productivity: A Data Science Approach

Hannah Kemper
University of Bonn, Germany

Abstract

Drought is among the most common but least understood phenomena that affect an increasing number of people in the context of climate change. To understand underlying drought dynamics affecting the local agricultural production in Botswana, a broad database comprising climatic and remote-sensing data together with socioeconomic indicators was set up. A data science approach that includes statistical and machine learning methods was chosen to retrieve information applicable in a drought early-warning system. The aim of the study was to examine how data science can contribute to the understanding of drought risk through the integration of various data sources. Different regression models (including linear and OLS) were applied. Naïve Bayes classification and Random Forest regression were included, as was a change point analysis. The impacts of two variables in particular, the Standardized Precipitation Index (SPI) and the Southern Oscillation Index (SOI), on crop productivity could be observed, highlighting possible national and regional thresholds. Further development of the early warning system, including validation, should be accompanied by ground-truth information and work with local partners.

Keywords:

data science, machine learning, agricultural production, drought early-warning, remote sensing

1 Introduction

Drought is a natural hazard characterized as 'a significant decrease in water availability during a prolonged period of time over a large area' (Keyantash & Dracup, 2002). If a drought occurs, the water balance turns negative and further impacts for systems relying on water should be expected. Immediate effects include high temperatures, high winds, and low relative humidity, as well as lower availability of surface and groundwater (Juana, 2014; Mishra & Singh, 2010; Wilhite, 2000). Further, there are significant impacts on natural, economic and social systems, including desertification and land degradation (Masih, Maskey, Mussá, & Trambauer, 2014). Droughts are considered among the most costly natural hazards, known to affect more people than any other hazard (AghaKouchak, 2015). Understanding the dynamics of drought impacts and the knock-on effects on local natural and human systems is still a challenging task (Bachmair, Kohn, & Stahl, 2015): drought remains one of the least understood natural hazards (Wilhite, 2000). Central in this study is 'agricultural drought': a lack of rain and diminishing soil moisture, resulting in crop loss. It is the direct consequence of persistent meteorological anomalies of dryness over significant periods of time within the agricultural cycle of a region.

Ambitious efforts exist to monitor dryness from space using drought indices or meteorological stations. Unfortunately, the long-term prediction of droughts remains a challenge, as the underlying dynamics of droughts are region-specific and important variables are only available in real-time. A Drought Early Warning System (DEWS) monitors and forecasts changes in temperature, precipitation, soil moisture and water bodies at the same time (World Meteorological Organization, 2006). A DEWS should integrate a wide range of indicators such as in-situ data (Bachmair, Stahl, et al., 2016), be comprehensive for immediate operationalization (Jain & Ormsbee, 2001), and consider the simultaneous occurrence of different impacts in different regions of the country (Wilhite, 2000). Remote sensing data and data science methods can be powerful tools to understand drought from statistical and environmental perspectives, and hence are used in this study.

This paper develops a workflow to analyse drought dynamics; it also identifies relevant data sources to support the development of a DEWS for Botswana. The overall aim of the study is to provide local authorities with new and important information on the phenomenon, overcoming the shortcomings of common solutions.

2 Research Area

The research area is semi-arid Botswana (see Figure 1), where 80% of the population is engaged in rain-fed agriculture and is consequently highly dependent on precipitation (Byakatonda, Parida, Kenabatho, & Moalafhi, 2019). The mostly flat topography is dominated by the Kalahari Desert, tropical grasslands and savannas. The rainfall occurs mainly during the austral summers (November to January) (Batisani & Yarnal, 2010). Botswana has suffered from frequent droughts in recent decades, especially from 1981–1987, 1991–1999, 2001–2005, 2007–2008, 2009–2010, 2010–2011, 2012–2013, 2014–2015, 2015–2016 and 2017–2019 (Statistics Botswana, 2020b). The most vulnerable groups are herdsmen, female-headed households, and low-income groups living in rural and remote areas (Fako & Molamu, 1995; Mugari, Masundire, & Bolaane, 2020). Although public awareness of drought risk is high (Akinyemi, 2017) and the government's efforts to import food have had positive effects on food security in the country, the vulnerability to droughts has not yet been reduced (Thinkhazard, 2020). Currently, the only sources of relevant information are a governmental drought monitoring system based on rainfall data, and a monthly meteorological bulletin. No information is available to anticipate developments over periods of several months (Department of Meteorological Service Agro-met Office).

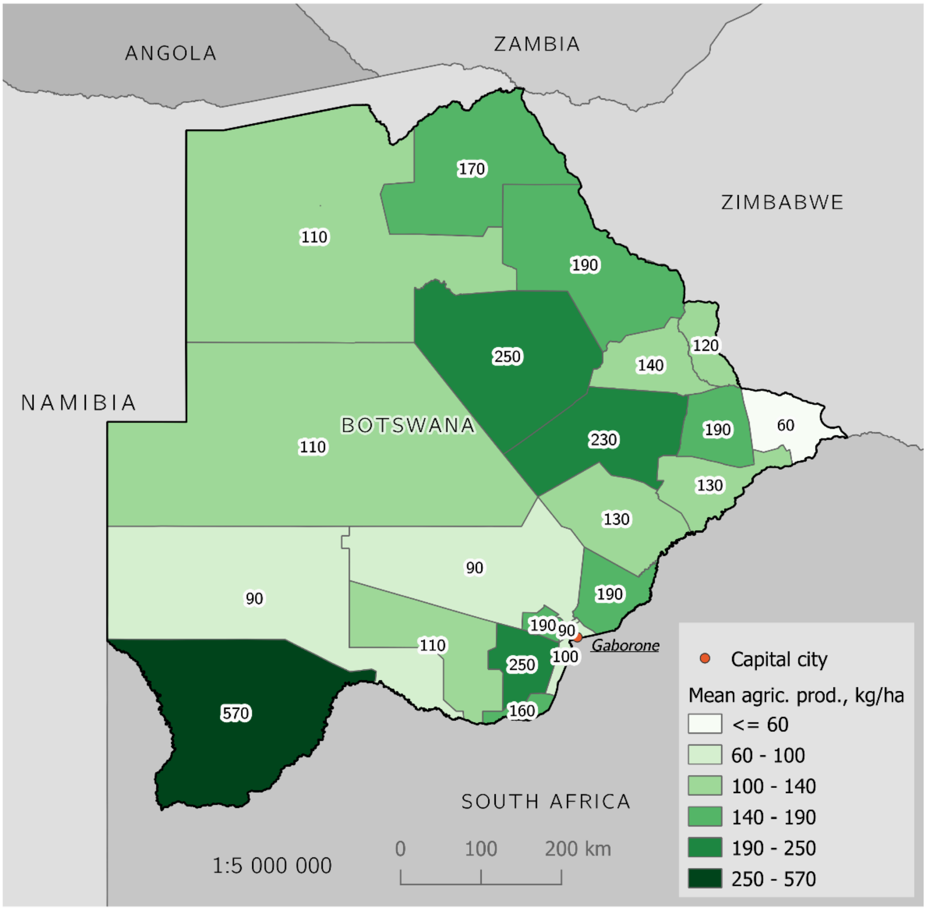

Figure 1: Map of research area showing the mean production of wheat, sorghum, millet and pulses in kg/ha.

3 Methodology

Data science generally refers to the application of versatile, both quantitative and qualitative, statistical methods to solve a problem. Machine learning (ML) is one of the most important techniques for predicting outcomes (Waller & Fawcett, 2013) as it overcomes the problems of traditional methods for handling huge amounts of data (Reichstein et al., 2019). Data Science approaches are iterative and must be repeated whenever research questions are modified, or new data is introduced. The analysis presented in this paper was conceptualized to examine three pillars of a DEWS: understanding, anticipating, and operationalizing actions to cope with drought risk (see Figure 2). To successfully mitigate and reduce the impact of droughts, a better understanding of local characteristics is needed.

A wide variety of data sources were combined to approach the research question broadly. The study period was from 1985 to 2020, while the research area included all agricultural districts of Botswana (see Figure 1).

The analysis was performed using the cloud-computing platform Google Earth Engine (GEE) and a script in Python 3.8.

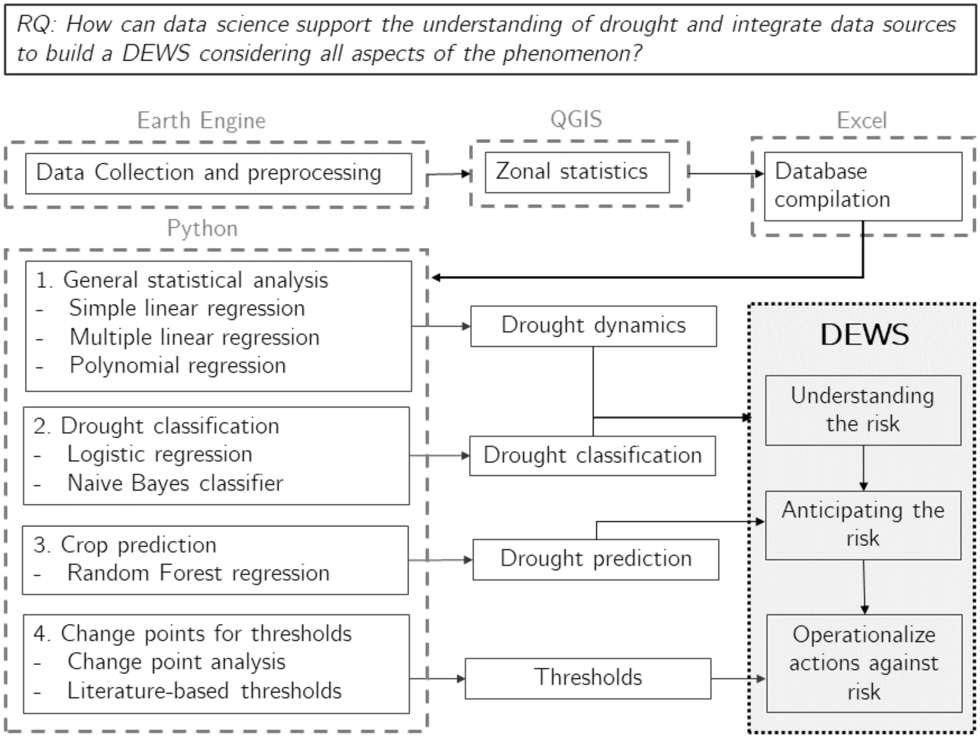

Figure 2: Contribution of Data Science methods to Drought Early Warning System (DEWS)

3.1 Data sources

A variety of datasets on climatic and vegetation conditions were combined with economic information on the agricultural production in the research area (see Table 1). The choice of variables was based on their presence in the research literature and their availability for the research area (Bachmair, Stahl, et al., 2016; Mishra & Singh, 2011).

Important seasonal changes in the precipitation regime were taken into account by creating a long-term and short-term variable for each indicator. 'Long-term' refers to the average values over 12 months from January to December; 'short-term' refers to the months from December to February. The sowing season in Botswana is limited to the summer rainy season (Food and Agriculture Organization of the United Nations, 2020). November and December are the most important months for crop planting, while December, January and February are the most important months for crop growth (Maruatona & Moses, 2021; Mugari et al., 2020).

Missing values were imputed using mean values; outliers of the 1.5 interquartile range were removed in order to obtain a more homogeneous data structure that permits easier regression analysis, and in order to reveal more relevant information (Wong & Wang, 2003). The shape of the dataset after the first cleansing was 40 variables and 791 rows. As some variables had large value differences, a standard scaler was applied to normalize the dataset in order to ensure that the models behaved well (Morid, Smakhtin, & Bagherzadeh, 2007).

Table 1: Overview of variables used in the study

variable	source
Crop production kg/ha	Statistics Botswana, 2020a
Drought period	Em-dat, C. R. E. D., 2010
Imports	Food and Agriculture Organization of the United Nations, 2021
Temperature	based on ERA-5 by Copernicus Climate Change Service, 2019
Precipitation	based on Chirps by Fick & Hijmans, 2017
Wind Speed	based on The Global Land Data Assimilation Project by Rodell et al., 2004
Southern Oscillation Index (SOI)	National Oceanic and Atmospheric Administration, 2021
North Atlantic Oscillation Index (NAOI)	National Oceanic and Atmospheric Administration, 2021
Palmer Drought Severity Index (PDSI)	based on TerraClimate by Abatzoglou, Dobrowski, Parks, & Hegewisch, 2018
Temperature Condition Index (TCI)	based on Temperature
Standardized Precipitation Index (SPI)	Funk et al., 2015
Normalized Differential Vegetation Index (NDVI)	based on Landsat 5, 7 & 8 by U.S. Geological Survey & NASA, 2021
Normalized Differential Water Index (NDWI)	based on Landsat 5, 7 & 8 by U.S. Geological Survey & NASA, 2021
Enhanced Vegetation Index (EVI)	based on Landsat 5, 7 & 8 by U.S. Geological Survey & NASA, 2021
Vegetation Condition Index (VCI)	based on NDVI
Vegetation Health Index (VHI)	based on VHI & TCI after Aksoy, Gorucu, & Sertel, 2019
Soil Moisture	based on The Global Land Data Assimilation Project by Rodell et al., 2004

3.2 Data analysis

Understanding the risk

A simple linear regression model was calculated for all variables in the dataset and presented in a correlation matrix (Figure 3). An Ordinary Least Squares model (OLS) was employed as a multiple linear regression (Pohlman & Leitner, 2003). For the evaluation of the model, R^2 and the Akaike information criterion (AIC) (Anderson & Burnham, 2002) were used. The Condition number (CN) (Dormann et al., 2013) and the Variance Inflation Factor (VIF) (Altman & Krzywinski, 2016) were used as checks for multicollinearity. Different combinations of variables were used, taking into account earlier results of the OLS and multiple linear regression models.

In order to reveal the possible dynamic nature of the variables, the polynomial regression was chosen as a non-linear approach (Ostertagová, 2012). It was conducted using degrees ranging from quadratic to higher-dimensional curves (Budescu, 1980).

Drought classification was realized using the information given in the Emdat database. The aim was to understand whether the variables differ substantially between periods of drought (marked 1) and non-drought (marked 0), and whether new data points could be classified correctly into the two categories.

Logistic Regression is ideal when handling dichotomous outcomes and has the advantage of being relatively simple to perform and interpret (Lever, Krzywinski, & Altman, 2016). Equation 1 describes a logistic regression (Sperandei, 2014), where π indicates the probability of an event and β_i are the regression coefficients with the reference group, and x_i is the explanatory variable.

Equation 1:

$$\log\left(\frac{\pi}{1-\pi}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_m x_m$$

The training data size was set to 70%, a common threshold (Dobbin & Simon, 2011). Mean squared error (MSE) and root-mean-square error (RMSE) were used as measures of the error size (Ostertagová, 2012), as they are very common in ML (Dormann et al., 2013).

The Naive Bayes (NB) classifier was chosen as another approach to classify the dataset. The NB follows the Bayes theorem, which takes outcome probabilities of related or dependent events into account by looking at conditional probabilities. Formula 2 (López Puga, Krzywinski, & Altman, 2015) indicates the posterior probability $P(A|B)$ using the prior probability of A, the probability of B, and the likelihood of a hypothesis of $P(B|A)$.

Equation 2:

$$P(A|B) = P(B|A) \times \frac{P(A)}{P(B)}$$

The Gaussian NB classifier was trained with 70% of the data points, and the accuracy score for the testing data of the model was calculated.

Anticipating the risk

A common ML algorithm applied for the prediction of numerical values is the random forest regression (RF). The RF is an ensemble of so-called regression trees in which several decision trees are combined (Strobl, Malley, & Tutz, 2009). Breiman (2001) suggests a formula that describes the random forest:

Equation 3:

$$m_{M,n}(X; \theta_1, \dots, \theta_m, \mathcal{D}_n) = \frac{1}{M}\sum_{j=1}^{M} m_n(x; \theta_j, \mathcal{D}_n)$$

where m_n is the predicted value, θ is a random variable, and D_n an independent variable. M represents the collection of trees fitted randomly with values in the dataset according to the input variables (Biau & Scornet, 2016).

For performance measures, the MSE, RMSE and R^2 were used (Bachmair, Svensson, Hannaford, Barker, & Stahl, 2016). Lastly, the percentage of correctly predicted values was calculated. A Randomized Search Cross-Validation was conducted with 3 folds on each of the following parameters: the number of trees, the depth of trees, the minimum samples per split, and the minimum samples per leaf. The result of this validation identified the best-performing parameters for the chosen independent variables (Koehrsen, 2018).

Operationalize against risk

There is no universal threshold of any indicator to identify the onset of a drought (Botterill & Hayes, 2012). Thresholds are not only specific to certain impact categories or affected sectors (Bachmair et al., 2015), but are also difficult to interpret when the underlying ecosystems are characterized by dynamic changes that follow the disequilibrium paradigm (see Skarpe, 1992). The following threshold concepts were considered for this work, based on Bachmair et al. (2015), Bachmair, Stahl, et al. (2016), and Chahuán-Jiménez, Rubilar, La Fuente-Mella, & Leiva (2021):

- median SPI values during drought periods of different agricultural districts
- median SOI and NAOI values during drought periods as long-term prediction indicators
- behaviour of variables around change points in crop yield data.

4 Results

4.1 Understanding the risk of drought impacts

The results of the linear regression are shown in Figure 3.

The overall performance of the OLS models was quite low regarding the R^2 values. The best fit of $R^2=0.193$ was achieved by a model using the SPI, SOI, Soil Moisture, NAOI, PDSI and TCI. It also scored better overall in the AIC. The CN and VIF values were always much lower than the threshold of 10 set by theory, indicating that there was no problem of multicollinearity (Salmerón, García, & García, 2018). The R^2 scores tended to rise with the number of variables but did not change substantially.

The accuracy scores of the polynomial regression ranged between -0.03 and +0.1. For the SPI_12, the highest score was achieved using a degree of 5. For the SOI_12, the highest value was attained using a degree of 2. Nevertheless, all accuracy scores showed low values (i.e. of less than 0.15). The overall performance of the Logistic Regression classifier showed accuracy values above 0.8, and MSE and RMSE values below 0.5. The PDSI and SPI in particular showed considerable differences between drought and non-drought periods. The average precipitation during droughts was roughly 25% lower than usual, and the temperature was slightly higher. The model using the TCI_12, SOI_12 and PRECIPITATION_12 variables had an accuracy of 0.96, and RMSE values of 0.2. Another model used the TCI_12 and SOI_12 variables and was evaluated as having an accuracy of 0.95. This result can be explained by the large differences between the drought and non-drought categories (see Appendix, Table 8).

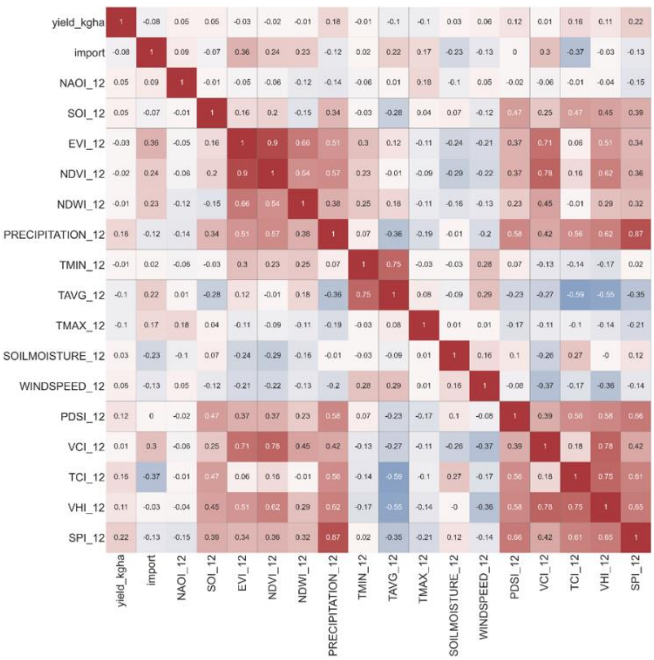

Figure 3: Correlation matrix

The NB classifier showed slightly lower accuracy values than the Logistic Regression. The highest-scoring models, with an accuracy of 0.89, used SOI both alone and in combination with rainfall data and NAOI. Accordingly, 89% of the testing data was classified correctly into drought and non-drought periods.

Table 1: Results of NB classifier & logistic regression

model	variables		logistic regression			NB
	dependent	independent	accuracy	MSE	RMSE	accuracy
1	drought_emdat	SOI_3, SOI_12	0.91	0.092	0.30	0.89
2	drought_emdat	SOI_12, PRECIPITATION_3, NAOI_12	0.89	0.105	0.324	0.89
3	drought_emdat	SOI_12, PRECIPITATION_3, TCI_12	0.92	0.080	0.283	0.84
4	drought_emdat	PRECIPITATION_12, yield_kgha, SOI_12, TCI_12	0.92	0.084	0.290	0.81
5	drought_emdat	TCI_12, SOI_12	0.95	0.046	0.215	0.88
6	drought_emdat	TCI_12, SOI_12, PRECIPITATION_12	0.96	0.042	0.205	0.87

4.2 Anticipating the risk of drought

Table 3 summarizes the results of the random forest regression. All models using the parameters derived from the Randomized Search Cross-Validation performed slightly better than the default model. However, the accuracy values range on a lower level, between 24% and 26%. The R^2 values range between 0.3 and 0.36.

Table 3: Results of random forest regression

Model	Variables	RMSE	R^2	Accuracy
default	all	0.246	0.34	24.6
n_estimators = 1800, max_depth = 90, max_features = 'sqrt', bootstrap = True. min_samples_split = 2, min_samples_leaf = 4	all	0.251	0.31	25.1
default	SOI_12, TMIN_3, TAVG_3, EVI_12	0.247	0.34	24.72
bootstrap=False, max_depth=10, max_features='sqrt', min_samples_leaf=2, min_samples_split=5, n_estimators=1200	SOI_12, TMIN_3, TAVG_3, EVI_12	0.2493	0.32	24.93
default	SPI_12	0.241	0.36	24.19
max_depth=50,	SPI_12	0.251	0.31	25.14

max_features='sqrt', min_samples_leaf=4, min_samples_split=10, n_estimators=800				
default	NDWI_12	0.243	0.35	24.38
default	SPI_12, SOI_12, PDSI_12, NAOI_12	0.246	0.34	24.61
n_estimators: 400, min_samples_split: 10, min_samples_leaf: 4, max_features: 'sqrt', max_depth: 90, bootstrap: True	SPI_12, SOI_12, PDSI_12, NAOI_12	0.253	0.3	25.3

4.3 Operationalize against drought risk

The median values of the SPI_12 variable during drought periods showed negative values ranging from -0.23 to -0.61. Lower SPI values were found in the surrounding districts, and the highest values (around -0.3) were found in the east and southeast of Botswana. The lowest value was found for Ngamiland district. The districts with the highest median values during drought periods were Bamalete-Tlokweng, Palapye, Bobonong and Barolong. Lower values were found in the northwest and higher values in the southwest. The values were lower than in non-drought conditions.

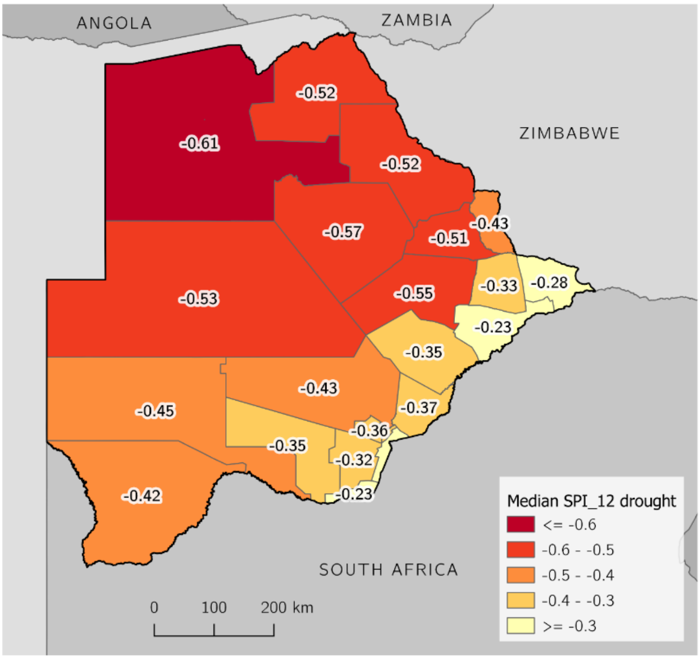

Figure 4: SPI_12 median values in drought years

As a second step to finding thresholds, the median SOI and NAOI values during drought periods were identified, as they were the only long-term prediction indicators in the database. The NAOI_3 was 0.48 during drought conditions and 0.645 in non-drought conditions. The NAOI_12 was 0.15 during droughts and 0.12 during normal conditions. The SOI shifted between negative and positive values. While the SOI_12 values were positive during normal conditions, the values dropped from 0.2 to -0.2 and -0.65 during drought conditions. Negative SOI values are associated with the onset of El Niño, and for this reason SOI values should be given high importance in establishing the DEWS.

Table 4: Thresholds derived from literature research

variable	drought onset	drought cessation
thresholds derived from median values of past drought events		
SOI_3	≤ 0	> 0
SOI_12	≤ 0	> 0
thresholds derived from median regional values of past drought events		
SPI_12	≤ -0.5 for Northwest ≤ -0.2 for Southeast ≤ -0.3 for all other areas	> -0.5 for Northwest > -0.2 for Southwest > -0.3 for all other areas
SPI_3	≤ -1.4 for Barolong & Ngwaketse S. ≤ -1.2 for Northwest and Southeast (except Barolong & Ngwaketse S.) ≤ -1.0 for Southwest and Centre ≤ -0.8 for East	> -1.4 for Barolong & Ngwaksetse S. > -1.2 for Northwest and Southeast (except Barolong & Ngwaketse S.) > -1.0 for Southwest and Centre > -0.8 for East

5 Discussion

Being a broad, flexible and globally applicable approach, the proposed workflow presents a wide range of statistical and ML methods that support the development of a DEWS for Botswana. Determining the most important variables influencing crop production in Botswana and further investigating their relationships and possible thresholds support an improved understanding of drought risk. This understanding can be used to monitor key variables and report important trend changes to the public. The approach presented here using Spatial Data Science for Early Warning is innovative, as scholars have previously focused, rather, on the prediction of indicators (see Chakrabarti, Bongiovanni, Judge, Zotarelli, & Bayer, 2014; Elliott, 2013; Kogan, Guo, & Yang, 2019; Potop, 2011). There are, however, some general restrictions on data quality and availability. Dividing the research area into smaller units or using a pixel-based approach could further enhance the precision of the analysis, as could including spatial data for the exact crop areas, if available.

The statistical analysis chosen was appropriate to the study case, using tried-and-tested, reliable methods. Regarding the regression analysis, the results showed surprisingly low correlations. This hints at a more complex relationship between the variables, or an issue with the quality of the data for agricultural production. Using the logistic regression and NB classifier was successful in both cases. Rather low accuracy values for the Random Forest indicated that a numerical prediction of the dependent variable was challenging in the heterogeneous dataset, yet low error measures demonstrated that the prediction was generally close to the value. The thresholds derived from change point analysis showed reasonable values in relation to other research findings that highlighted differences between regions. Including variables like the NAOI and SOI that can be forecast is highly to be recommended for a DEWS. Because of global trends like climate change, these thresholds should be verified and updated in the future. This is necessary as the rising temperature will affect all other variables, and thresholds that are reliable now may no longer be so in the future.

Another significant shortcoming lies in the absence of ground-truth data. Therefore, the investigation of local coping strategies, the calendar shift of the analysis months to austral summers, and the validation of the proposed workflow with rural and even indigenous communities are potentials that could be explored in the future. Further, the disaster context was very specifically focused on droughts. Multi-hazards or cascading effects should be considered in subsequent studies (see Gill & Malamud, 2016; Pescaroli & Alexander, 2018).

6 Conclusion

A methodology using different statistical and ML methods following a data science approach was applied to the case of a DEWS for Botswana. Droughts being a highly relevant topic for local agriculture, important findings were made, using several globally available datasets, regarding the negative effects on crop yield. The most important threshold for drought onset is 0 for the SOI, which could be used in combination with the SPI. Ground truth verification and validation should be envisioned for future developments of the DEWS in Botswana. To be highlighted is the applicability, in different research areas, of this methodology regarding the identification of thresholds.

Acknowledgements

I would like to thank Prof. Dr. Greve (University Bonn) and Dr. Felicia Akinyemi (University Bern) for so thoroughly supporting my research on Botswana. Further, I would like to express my gratitude towards Sebastian Schmidt (University Salzburg) and Theresa Keller (BfG) for their helpful comments in all stages of my research.

References

Abatzoglou, J. T., Dobrowski, S. Z., Parks, S. A., & Hegewisch, K. C. (2018). Terraclimate, a high-resolution global dataset of monthly climate and climatic water balance from 1958-2015. *Scientific Data*, *5*(1), 170191. https://doi.org/10.1038/sdata.2017.191

AghaKouchak, A. (2015). A multivariate approach for persistence-based drought prediction: Application to the 2010–2011 East Africa drought. *Journal of Hydrology*, *526*, 127–135. https://doi.org/10.1016/j.jhydrol.2014.09.063

Akinyemi, F. O. (2017). Climate Change and Variability in Semiarid Palapye, Eastern Botswana: An Assessment from Smallholder Farmers' Perspective. *Weather, Climate, and Society*, *9*(3), 349–365. https://doi.org/10.1175/WCAS-D-16-0040.1

Aksoy, S., Gorucu, O., & Sertel, E. (2019). Drought Monitoring using MODIS derived indices and Google Earth Engine Platform. In *2019 8th International Conference* (pp. 1–6). https://doi.org/10.1109/Agro-Geoinformatics.2019.8820209

Altman, N., & Krzywinski, M. (2016). Regression diagnostics. *Nature Methods*, *13*(5), 385–386. https://doi.org/10.1038/nmeth.3854

Anderson, D. R., & Burnham, K. P. (2002). Avoiding Pitfalls When Using Information-Theoretic Methods. *The Journal of Wildlife Management*, *66*(3), 912–918.

Bachmair, S., Kohn, I., & Stahl, K. (2015). Exploring the link between drought indicators and impacts. *Natural Hazards and Earth System Sciences*, *15*(6), 1381–1397. https://doi.org/10.5194/nhess-15-1381-2015

Bachmair, S., Stahl, K., Collins, K., Hannaford, J., Acreman, M., Svoboda, M., . . . Overton, I. C. (2016). Drought indicators revisited: the need for a wider consideration of environment and society. *Wiley Interdisciplinary Reviews: Water*, *3*(4), 516–536. https://doi.org/10.1002/wat2.1154

Bachmair, S., Svensson, C., Hannaford, J., Barker, L. J., & Stahl, K. (2016). A quantitative analysis to objectively appraise drought indicators and model drought impacts. *Hydrology and Earth System Sciences*, *20*(7), 2589–2609. https://doi.org/10.5194/hess-20-2589-2016

Batisani, N., & Yarnal, B. (2010). Rainfall variability and trends in semi-arid Botswana: Implications for climate change adaptation policy. *Applied Geography*, *30*(4), 483–489. https://doi.org/10.1016/j.apgeog.2009.10.007

Biau, G., & Scornet, E. (2016). A random forest guided tour. *TEST*, *25*(2), 197–227. https://doi.org/10.1007/s11749-016-0481-7

Botterill, L. C., & Hayes, M. J. (2012). Drought triggers and declarations: science and policy considerations for drought risk management. *Natural Hazards*, *64*(1), 139–151. https://doi.org/10.1007/s11069-012-0231-4

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*, 5–32.

Budescu, D. V. (1980). A Note On Polynomial Regression. *Multivariate Behavioral Research*, *15*(4), 497–506. https://doi.org/10.1207/s15327906mbr1504_7

Byakatonda, J., Parida, B. P., Kenabatho, P. K., & Moalafhi, D. B. (2019). Prediction of onset and cessation of austral summer rainfall and dry spell frequency analysis in semiarid Botswana. *Theoretical and Applied Climatology*, *135*(1-2), 101–117. https://doi.org/10.1007/s00704-017-2358-4

Chahuán-Jiménez, K., Rubilar, R., La Fuente-Mella, H. de, & Leiva, V. (2021). Breakpoint Analysis for the COVID-19 Pandemic and Its Effect on the Stock Markets. *Entropy (Basel, Switzerland)*, *23*(1). https://doi.org/10.3390/e23010100

Chakrabarti, S., Bongiovanni, T., Judge, J., Zotarelli, L., & Bayer, C. (2014). Assimilation of SMOS Soil Moisture for Quantifying Drought Impacts on Crop Yield in Agricultural Regions. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, *7*(9), 3867–3879. https://doi.org/10.1109/JSTARS.2014.2315999

Copernicus Climate Change Service (2019). *ERA5-Land monthly averaged data from 2001 to present*. ECMWF. https://doi.org/10.24381/CDS.68D2BB30

Department of Meteorological Service Agro-met Office (July 2021). *Botswana Agrometeorological Monthly Bulletin* [Press release]. Gaborone.

Dobbin, K. K., & Simon, R. M. (2011). splitting training and testing data. *Medical Genomics*, *4*(31).

Dormann, C. F., Elith, J., Bacher, S., Buchmann, C., Carl, G., Carré, G., . . . Lautenbach, S. (2013). Collinearity: a review of methods to deal with it and a simulation study evaluating their performance. *Ecography*, *36*(1), 27–46. https://doi.org/10.1111/j.1600-0587.2012.07348.x

Elliott, J. (2013). *Simulated county- and state-level maize yields, 1979-2012*. figshare. https://doi.org/10.6084/M9.FIGSHARE.501263

Em-dat, C. R. E. D. (2010). *The OFDA/CRED international disaster database*. Retrieved from www.emdat.be

Fako, T. T., & Molamu, L. (1995). The Seven-Year Drought, Household Food Security and Vulnerable Groups in Botswana. *Botswana Journal of African Studies*, *9*(2), 48–70.

Fick, S. E., & Hijmans, R. J. (2017). WorldClim 2: new 1-km spatial resolution climate surfaces for global land areas. *International Journal of Climatology*, *37*(12), 4302–4315. https://doi.org/10.1002/joc.5086

Food and Agriculture Organization of the United Nations (2020). GIEWS Country Brief Botswana.

Food and Agriculture Organization of the United Nations (2021). *FAOSTAT*. Retrieved from http://www.fao.org/faostat/en/#data

Funk, C., Peterson, P., Landsfeld, M., Pedreros, D., Verdin, J., Shukla, S., . . . Michaelsen, J. (2015). The climate hazards infrared precipitation with stations--a new environmental record for monitoring extremes. *Scientific Data*, *2*(1), 150066. https://doi.org/10.1038/sdata.2015.66

Gill, J. C., & Malamud, B. D. (2016). Hazard interactions and interaction networks (cascades) within multi-hazard methodologies. *Earth System Dynamics*, *7*(3), 659–679. https://doi.org/10.5194/esd-7-659-2016

Jain, A., & Ormsbee, L. (2001). A decision support system for drought characterization and management. *Civil Engineering and Environmental Systems*, *18*(2), 105–140. https://doi.org/10.1080/02630250108970296

Juana, J. (2014). Socioeconomic Impact of Drought in Botswana. *International Journal of Environment and Sustainable Development*, *11*(1), 43–60.

Keyantash, J., & Dracup, J. A. (2002). The Quantification of Drought: An Evaluation of Drought Indices. *Bulletin of the American Meteorological Society*, 1167–1180.

Koehrsen, W. (2018, January 10). Hyperparameter Tuning the Random Forest in Python. *Towardsdatascience*. Retrieved from https://towardsdatascience.com/hyperparameter-tuning-the-random-forest-in-python-using-scikit-learn-28d2aa77dd74

Kogan, F., Guo, W., & Yang, W. (2019). Drought and food security prediction from NOAA new generation of operational satellites. *Geomatics, Natural Hazards and Risk*, *10*(1), 651–666. https://doi.org/10.1080/19475705.2018.1541257

Lever, J., Krzywinski, M., & Altman, N. (2016). Logistic regression. *Nature Methods*, *13*(7), 541–542. https://doi.org/10.1038/nmeth.3904

López Puga, J., Krzywinski, M., & Altman, N. (2015). Points of significance: Bayes' theorem. *Nature Methods*, *12*(4), 277–278. https://doi.org/10.1038/nmeth.3335

Maruatona, P. B., & Moses, O. (2021). Assessment of the onset, cessation, and duration of rainfall season over Botswana. *Modeling Earth Systems and Environment*, 1–12. https://doi.org/10.1007/s40808-021-01178-5

Masih, I., Maskey, S., Mussá, F. E. F., & Trambauer, P. (2014). A review of droughts on the African continent: a geospatial and long-term perspective. *Hydrology and Earth System Sciences*, *18*(9), 3635–3649. https://doi.org/10.5194/hess-18-3635-2014

Mishra, A. K., & Singh, V. P. (2010). A review of drought concepts. *Journal of Hydrology*, *391*(1-2), 202–216. https://doi.org/10.1016/j.jhydrol.2010.07.012

Mishra, A. K., & Singh, V. P. (2011). Drought modeling – A review. *Journal of Hydrology*, *403*(1-2), 157–175. https://doi.org/10.1016/j.jhydrol.2011.03.049

Morid, S., Smakhtin, V., & Bagherzadeh, K. (2007). Drought forecasting using artificial neural networks and time series of drought indices. *International Journal of Climatology*, *27*(15), 2103–2111. https://doi.org/10.1002/joc.1498

Mugari, E., Masundire, H., & Bolaane, M. (2020). Effects of Droughts on Vegetation Condition and Ecosystem Service Delivery in Data-Poor Areas: A Case of Bobirwa Sub-District, Limpopo Basin and Botswana. *Sustainability*, *12*(19), 8185. https://doi.org/10.3390/su12198185

National Oceanic and Atmospheric Administration (2021). *Teleconnections*. Retrieved from https://www.ncdc.noaa.gov/teleconnections/

Ostertagová, E. (2012). Modelling using Polynomial Regression. *Procedia Engineering*, *48*, 500–506. https://doi.org/10.1016/j.proeng.2012.09.545

Pescaroli, G., & Alexander, D. (2018). Understanding Compound, Interconnected, Interacting, and Cascading Risks: A Holistic Framework. *Risk Analysis : An Official Publication of the Society for Risk Analysis*, *38*(11), 2245–2257. https://doi.org/10.1111/risa.13128

Pohlman, J., & Leitner, D. (2003). A Comparison of Ordinary Least Squares and Logistic Regression. *Ohio Science Journal*, *103*(5), 118–125.

Potop, V. (2011). Evolution of drought severity and its impact on corn in the Republic of Moldova. *Theoretical and Applied Climatology*, *105*(3-4), 469–483. https://doi.org/10.1007/s00704-011-0403-2

Reichstein, M., Camps-Valls, G., Stevens, B., Jung, M., Denzler, J., Carvalhais, N., & Prabhat (2019). Deep learning and process understanding for data-driven Earth system science. *Nature*, *566*(7743), 195–204. https://doi.org/10.1038/s41586-019-0912-1

Rodell, M., Houser, P. R., Jambor, U., Gottschalck, J., Mitchell, K., Meng, C., . . . Toll, D. (2004). The Global Land Data Assimilation System. *Bulletin of the American Meteorological Society*, *85*(3), 381–394. https://doi.org/10.1175/BAMS-85-3-381

Salmerón, R., García, C. B., & García, J. (2018). Variance Inflation Factor and Condition Number in multiple linear regression. *Journal of Statistical Computation and Simulation*, *88*(12), 2365–2384. https://doi.org/10.1080/00949655.2018.1463376

Skarpe, C. (1992). Dynamics of savanna ecosystems. *Journal of Vegetation Science*, *3*, 293–300.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia Medica*, *24*(1), 12–18. https://doi.org/10.11613/BM.2014.003

Statistics Botswana (2020a). *Annual Agricultural Survey Report 2019: Traditional Sector*. Gaborone.

Statistics Botswana (2020b). *Botswana Environment Statistics Natural and Technological Disasters Digest.*

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323–348. https://doi.org/10.1037/a0016973

Thinkhazard (2020). Botswana. Retrieved from https://thinkhazard.org/en/report/35-botswana/DG

U.S. Geological Survey, & NASA (2021). *Landsat Collections in Earth Engine: Collection 1: Landsat8, Landsat 7 & Landsat 5*. Retrieved from https://developers.google.com/earth-engine/datasets/catalog/landsat

Waller, M. A., & Fawcett, S. E. (2013). Data Science, Predictive Analytics, and Big Data: A Revolution that will transform supply chain design and management. *Journal of Business Logistics*, *34*(2), 77–84.

Wilhite, D. A. (2000). *Drought: A Global Assessment*. London: Routledge.

Wong, A., & Wang, Y. (2003). Pattern discovery: a data driven approach to decision support. *IEEE Transactions on Systems, Man and Cybernetics*, *33*(1), 114–124. https://doi.org/10.1109/TSMCC.2003.809869

World Meteorological Organization (2006). Drought monitoring and early warning: concepts, progress and future challenges. Retrieved from www.wamis.org/agm/pubs/brochures/ WMO1006e.pdf

Appendix

Table 5: Overview of variables

variable	mean	minimum	median	maximum
yield_kgha	172.71	0	92.49	7,723.92
import	11,1287.71	0	90720	246,657
NAOI_3	0.42	-1.67	0.56	1.66
NAOI_12	0.04	-1.9	0.13	2.63
SOI_3	0.06	-1.9	0.13	2.63
SOI_12	0.08	-0.93	-0.02	1.4
EVI_3	0.28	0.09	0.28	0.50
EVI_12	0.21	0.09	0.20	0.35
NDVI_3	0.28	0.09	0.28	0.47
NDVI_12	0.23	0.11	0.23	0.4
NDWI_3	-0.01	-0.19	-0.01	0.24
NDWI_12	-0.08	-0.21	-0.09	0.053
PRECIPITATION_3	240.37	41.98	216.45	698.44
PRECIPITATION_12	406.94	133.73	392.93	783.06
TMIN_3	19.23	16.38	19.27	21.37
TMIN_12	13.86	10.59	13.76	17.08
TAVG_3	26.07	22.10	26.08	30.72
TAVG_12	22.32	18.78	22.34	25.63
TMAX_3	32.55	28.51	32.55	37.08
TMAX_12	49.39	26.33	30.17	387.63
SOILMOISTURE_3	19.68	4.48	18.84	77.6
SOILMOISTURE_12	16.19	4.71	16.19	31.97
WINDSPEED_3	5.51	3.97	5.45	7.35
WINDSPEED_12	5.82	4.81	5.83	6.99
PDSI_3	-30.23	-518.94	-30.27	649.73
PDSI_12	-47.38	-434.63	-81.13	950.01
VCI_3	0.5	0	0.5	1
VCI_12	0.53	0	0.53	1
TCI_3	0.5	0	0.49	1
TCI_12	0.5	0	0.49	1
VHI_3	0.5	0	0.50	0.99
VHI_12	0.51	0.08	0.51	0.99
SPI_3	-0.43	-1.93	-0.63	2.88
SPI_12	0	-1.09	-0.07	1.63

Table 6: OLS performance

No.	r^2	adj. R^2	AIC	CN	VIF	
1	0.105	0.103	249.9	2.58	SOI_3 SPI_12	1.02
2	0.133	0.131	225.2	2.9	SOI_12 SPI_12	1.16
3	0.079	0.076	273.1	3.31	PRECIPITATION_3 VHI_3	1.23
4	0.111	0.108	245.2	3.69	PDSI_12 SPI_12	1.46
5	0.108	0.105	248	3.49	VHI_12 SPI_12	1.49
6	0.105	0.102	250.5	3.18	TCI_12 SPI_12	1.33
7	0.134	0.131	226.3	3.45	SPI_12 SOI_12 TCI_12	1.39 1.26 1.45
8	0.133	0.130	227.1	3.79	SPI_12 SOI_12 VHI_12	1.54 1.26 1.61
9	0.192	0.188	173.7	4.25	SPI_12 SOI_12 PDSI_12 NAOI_12	1.59 1.27 1.55 1.07
10	0.189	0.185	176.4	3.82	SPI_12 SOI_12 TCI_12 NAOI_12	1.49 1.29 1.46 1.08
11	0.155	0.150	209	4.34	SPI_12 TCI_12 PDSI_12 NAOI_12	1.768 1.40 1.51 1.05
12	0.192	0.187	175.6	4.64	SPI_12 SOI_12 TCI_12 NAOI_12 PDSI_12	1.80 1.34 1.48 1.08 1.57
13	0.192	0.187	175.3	4.57	SPI_12 SOI_12 VHI_12 NAOI_12	1.83 1.32 1.78 1.08

					PDSI_12	1.69
14	0.192	0.186	176.9	5	SPI_12 SOI_12 VHI_12 NAOI_12 PDSI_12 TCI_12	1.91 1.36 2.09 1.08 1.69 1.74
15	0.193	0.187	176.4	4.81	SPI_12 SOI_12 SOILMOISTURE_12 NAOI_12 PDSI_12 TCI_12	1.81 1.35 1.22 1.1 1.59 1.75

Table 7: Polynomial regression

variable	degrees	score	variable	degrees	score
SPI_12	2 3 4 5 6	0.102308 0.098425 0.104660 0.103773 0.102503	NAOI_12	2 3 4 5 6	0.014810 -0.00583 0.024961 0.027001 0.046439
SOI_12	2 3 4 5 6	-0.02483 -0.024737 -0.047782 -0.008321 -0.007743	VHI_12	2 3 4 5 6	0.048829 0.031246 0.027365 0.018343 0.018299
TCI_12	2 3 4 5 6	0.023017 0.063382 0.066476 0.070053 0.072128	TAVG_12	2 3 4 5 6	0.016366 0.007817 0.006015 -0.018565 -0.01859
PRECIPITATION_12	2 3 4 5 6	0.087858 0.088270 0.086982 0.089702 0.087151	NDVI_12	2 3 4 5 6	-0.030134 -0.030066 -0.031589 -0.031800 -0.061702

Table 8: Differences between drought and non-drought periods

variable	mean non-drought (Emdat)	mean drought (Emdat)	median non-drought (Emdat)	median drought (Emdat)
yield kg/ha	659.0	132.0	146.3	59.2
import	107,250	131,581	90,720	93,171
NAOI_12	0.0531	-0.0467	0.15	0.15
SOI_12	0.18968	-0.4683	0.17	-0.645
EVI_12	0.2	0.2	0.2	0.2
NDVI_12	0.23	0.22	0.23	0.23
NDWI_12	-0.086	-0.064	-0.08	-0.08
PRECIPITATION_12	425.03	316.61	412.6	318.87
TMIN_12	13.86	13.87	13.74	13.86
TAVG_12	22.16	23.1	22.16	23.17
SOILMOISTURE_12	16.4	15.16	16.19	15.11
WINDSPEED_12	5.82	5.88	5.82	5.86
PDSI_12	-9.72	-235.41	-47.4	-250.4
VCI_12	0.53	0.52	0.52	0.53
TCI_12	0.55	0.26	0.55	0.25
VHI_12	0.53	0.39	0.53	0.39
SPI_12	0.07	-0.42	0.02	-0.39