6th Intercontinental Geoinformation Days

igd.mersin.edu.tr

The effect of climatic factors on the cotton productivity using machine learning approaches

Bakhtiyar Babashli *1

[1]*National Aviation Academy, Faculty of Aerospace, Department of Aerospace Environmental Monitoring, Baku, Azerbaijan*

Keywords
Cotton
Machine learning
Climate
Factor
Productivity

Abstract

Agriculture and the farming are two of the most important sectors of the economy. An accurate and timely assessment of cotton field productivity is useful for management decisions about cotton supply and sales. Cotton production concentration is influenced by a variety of factors. The impact of climatic conditions (rainfall, temperature, wind, etc.) on cotton productivity is studied in order to determine the quantitative relationship between these parameters and productivity. Several machine learning techniques have been researched and used to estimate crop yield. Errors like RMSE, MSE, MAE, and R2 were employed as indicators, and the polynomial regression model was chosen as the best among them.

1. Introduction

Cotton is one of the most important and widely produced products in the world. Currently, cotton (Gossypium spp.) a natural fiber produced on a large scale is a source of income for millions of farmers. Cotton cultivation is mostly adapted to temperate, tropical and subtropical climates worldwide, but its future development may take a different path due to future climate change (Bange et al., 2016). Climate change can have an impact on cotton output in both good and negative ways. Temperature influences cotton growth and development by determining fruit production, photosynthesis, and respiration rates (Turner et al., 1986).

The conducted research showed that the achievements of science and advanced practice in cotton farming the productivity of cotton can be further increased by applying it (Seyidaliyev, 2012). It is required to enhance cotton production, quality, fast-growing cotton kinds, crop collecting, and the use of new technology.

Cotton yield is statistically significantly related to both water availability and temperature (Crane-Droesch, 2018). Both of these factors have a larger effect on the cotton yield than soil acidity. This is important because it shows that many factors have an effect on cotton growth, and we must consider all of these factors when manipulating the genetic makeup of the cotton plant.

Among the most extensively produced crops for the production of fiber globally, cotton is economically important and is grown in more than 60 nations with temperate and tropical climates (Jans et al, 2021). However, a rise in temperature during the cotton growth season (CGS) had a negative impact on yield (Snider et al, 2009). Furthermore, an increased frequency of hot days may be detrimental to cotton yield.

This project's objective is to forecast cotton yield using climate information in Azerbaijan. The influence of changing climate conditions on productivity was assessed using machine learning techniques. This problem is statistically based on a regression model. As a result, our research sought to identify any quantifiable association between environmental conditions and production. Particularly, several environmental elements appear to be related to fertility. The primary goal of this research is to put to the test the current statistical link between environmental conditions and productivity.

Modern machine learning algorithms were applied to the study to increase the model's superiority. This, it is possible to predict cotton productivity by learning machine algorithms to evaluate the effect of climate elements on cotton productivity in the cotton-producing regions of Azerbaijan.

2. Method and data

Data on the climate and cotton yields from 2017 to 2021 were used in the current analysis, including information on the effects of daily temperature, wind, humidity and rainfall on cotton yields. These parameters during the previous five years varied significantly,

* **Corresponding Author**

*(bakhtiyar.babashli@gmail.com) ORCID ID 0000-0001-7931-1677

Cite this study

Babashli B (2023). The Effect of Climatic Factors on the Cotton Productivity Using Machine Learning Approaches. Intercontinental Geoinformation Days (IGD), 6, 176-179, Baku, Azerbaijan

according to the results of daily data from 16 locations for each growing season.

2.1. Climate and crop yield data

The information is gathered from several public data sources. NASA is the source of the weather information (Nasa data). Productivity use the website of the Statistical Commission of the Republic of Azerbaijan done (The State Statistical Committee of the Republic of Azerbaijan).

The climate (temperature, rainfall, humidity, wind) and soil data (soil wetness and moisture) for the last five years are considered climate data. The information on the NASA website basically consisted of the daily. Beyleqan, Fuzuli, Agcabedi, Aghdam, Terter, Yevlax, Zerdab, Kurdemir, Imishli, Saatli, Bilesuvar, Neftchala, Salyan, Sabirabad, Goranboy and Berde were the 16 regions that were taken into consideration for the study (figure 1). According to the Ministry of Science and education of the Republic of Azerbaijan "cotton-growing" textbook guidelines for field crop planting and harvesting dates, the length of the climatological growth season (CGS) was counted from May to September (Pambıqçılıq, 2017). The daily weather and yield data were collected over growing seasons.

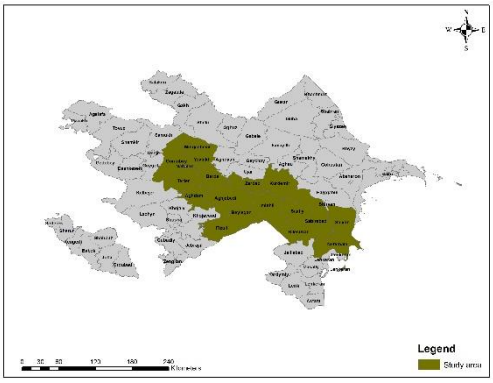

Figure 1. Study area

Using Matplotlib scatter plots were created (figure 2). The data in the first diagram in the figure 2 shows that the wind speed ranges from 0.51 to 9.47m/s. In the second of the diagrams the humidity is 13.69 and 89.94 varies between. In the third diagram shows the rainfall ranges from 0 to 62.85 mm/day. In the fourth diagram temperature ranges from 1.31 to 24.59℃. In the fifth diagram soil wetness ranges from 0.1 to 0.8. In the sixth diagram soil moisture ranges from 0.37 to 0.72. In the seventh diagram indicates surface pressure ranges from 91.83 to 82.95 kPa and the eighth diagram display cotton yield ranges from 8.2 to 40.8 cent/ha.

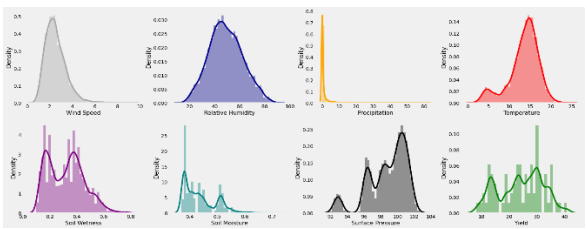

Figure 2. Distribution for climate variables

2.2. Method

In this study, trends were evaluated using the regression method. The fundamental concept of the approach is to establish an algebraic relationship between the dependent and independent variables. A model of the relationship and estimations of the parameter values are used to generate an estimated regression equation (Ostertagováa, 2012). The linear regression method was employed in this investigation. A statistical tool called linear regression can be used to determine the relationship between several "explanatory" factors and a real-valued outcome. This study use nonlinear polynomial predictors. A nonlinear polynomial predictor, also known as an n-dimensional one-dimensional polynomial function.

$$y = a0 + a1\,x + a2\,x\,2 + \bullet\bullet\bullet + an\,x\,n \quad (1)$$

where a vector of coefficients of size n + 1 is (a0, a1, a2, , an).

We have used various parameters as our independent variable (climate data) and yield value as our dependent variable to apply this methodology (Gonzalez-Sanchez et al, 2014). To distinguish between train and test data, the dichotomy of data was developed. The following variables are used in this experiment: y is the cotton yield; a0 is the intercept; a1, a2, a3... an, is the coefficient for rainfall, temperature, humidity and wind etc. The model for this experiment was built using the scikit-learn package's different regression function. Algorithms used for productivity prediction are as follows: Polynomial Regression (MPR), Random Forest Regressor (RFT), Extreme Gradient Boosting Regressor (XGBR), Cat Boost Regression (CBR), LGBM Regressor (LGBR).

Polynomial Regression is a form of regression analysis in which the relationship between independent variables and dependent variables is modeled in an nth-degree polynomial (Mohammad et al. 2022). Polynomial regression models are usually fit by the least squares method. The least squares method minimizes the variance of the coefficients according to the Gauss Markov theorem. One of the fundamental differences between Linear and Polynomial Regression is that Polynomial Regression does not require the relationship between the independent and dependent variables in the data set to be linear. Polynomial Regression is commonly employed when the Linear Regression Model fails to capture the points in the data and the Linear Regression doesn't succeed to properly describe the best outcome. Linear regression is essentially a first-degree polynomial regression.

Random Forest is a decision tree-based machine learning algorithm. Random Forest employs an ensemble approach, which is highly popular in recent years. Ensemble is a common decision-making system based on decision trees. Random Forest regressor is an approach for predicting vegetation that generates random trees (Breiman et al. 2001). The benefit of utilizing Random Forest allows us to eliminate difficulties caused by outliers in our database.

Extreme Gradient Boosting Regression is a machine learning technique for classification problems that

generates a set of prediction models, usually in the form of decision trees. The model is improved on the existing gradient boosting algorithm (Chen et al, 2016). This method combines a number of key factors to estimate complex statistical dependencies (Thomaset et al, 2018). This model is widely used because it gives positive forecast results. The main goal of XGB is to speed up and improve the performance of decision trees.

The most modern productivity boosting technique is the Cat Boost Regressor (CAT). This is due to the implementation of a more efficient gradient boosting tree algorithm (Khan et al, 2020). Categorical parameters, symmetric indicators with minimal variables and superior accuracy form the basis of this decision tree algorithmic framework. Like Gradient Boosting and XGB, the CatBoost method builds multiple decision trees at the same time each time it tries to reduce the error.

LightGBM is a histogram-based method. Learning decision trees can be done using one of the two approaches: level-wise (depth-wise) or leaf-wise (Shahhosseini et al, 2021). The equilibrium of the tree is maintained as the tree grows using the level-wise technique. The division process continues from the leaves in the leaf-wise technique, which lowers loss. LightGBM distinguishes itself from other boosting algorithms because to this property. The Leaf-wise technique reduces error rates and speeds up learning. However, the Leaf-wise growth technique causes the model to over-learn when the amount of data points is small. As a result, the technique is more suited for usage in huge data.

We are utilizing the coefficient of determination as the experiment's evaluation metric. The amount of variance in the dependent variable that can be explained by the predictors in the model is measured by the coefficient of determination R2, which is well - defined in linear regression models (Mohammed et al, 2022). R2 has a value between 0 and 1, with 0 being the worst and 1 being the greatest.

3. Results

This study used several regression models to predict climate data to increase cotton productivity and the results were compared. The accuracy of these forecasting models is measured by R Square, Root Mean Square Error (RMSE) and Mean Percent Prediction Error (MPPE).

Table 1. Accuracy Results of Algorithms

Model	MAE	MSE	RMSE	R Square
Polynomial	0.007	0.0001	0.009	0.960
XGBoost	0.012	0.0003	0.017	0.871
CatBoost	0.013	0.0003	0.017	0.867

Polynomial regression is the best model for our research, with R-squared values of 0.96, MAE of 0.0073, MSE of 0.0001, and RMSE of 0.0094 presented in table 1. This shows that the experiment's model was quite accurate. As a result, we discover that there is a quantitative association between environmental conditions and cotton output utilizing the information from the aforementioned numerous studies. The results of this experiment also show that it is possible to forecast cotton's return using a variety of environmental conditions.

4. Discussion

The relationship between mainly meteorological factors and cotton yield was analyzed by Polynomial Regression algorithm. The obtained results show that temperature, cloud cover, rain and wind are the most important factors affecting the productivity of cotton at all growth stages. So, timely action due to these factors can both increase the production and prevent the reduction of cotton productivity. Meteorological elements have different effects on cotton productivity. For example, increased rainfall, potential evapotranspiration and excess water during the cycle can lead to increased cotton yield. On the contrary, a decrease in cotton productivity is observed due to the increase in air temperature and water shortage (Silva et al. 2021). One of the main factors of climate change is the lack of water, which has a great impact on agriculture and reduces the productivity and quality of various crops.

Cotton yield potential is determined between planting and flowering dates. Climatic conditions during this period have a great impact on productivity. Climatic elements have shown different relationships and intensities with cotton productivity. Cotton grows most effective at temperatures between 25-30^{0}C. When the temperature drops below 25^{0}C, plant growth slows down. Raising the temperature to 300°C stimulates cotton development in the early stage of plant development, before budding. Above 35°C, the cotton plant's growth and development processes are halted. The seeds of the plant heat up more at this greater temperature and their development ceases.

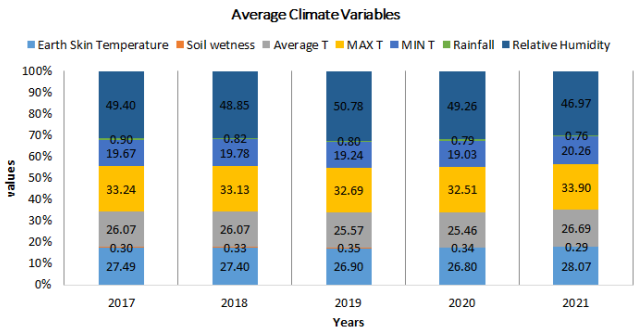

Figure 3. Average climate variables

Cotton yield in the researched area varied between 8.2 cents/ha in 2017 (Nefchala) and 40.8 cents/ha (Salyan) in 2020. İn 2017 and 2018, the productivity of cotton was observed to be lower than in other years. During these years, the average temperature was quite low, with the maximum average temperature exceeding 40^{0}C Celsius. Simultaneously, beginning in 2019, new technologies and the usage of various fertilizers and chemicals were implemented. Furthermore, as science has advanced, the sowing of new seeds has increased output. Nonetheless, while there are irrigation issues in 2020-2021, the introduction of new technology has produced circumstances for increased productivity.

The polynomial algorithm was the best performing algorithm to track all spatial variability of cotton productivity using climate elements as independent variables. As the temperature increases in the model, it tends to increase its performance. However, increasing the ranks of the model also increases the risk of overfitting and underfitting the data. Forward Selection and Backward Selection can be used to find the correct model rank to avoid excessive or inappropriate. These methods increase or decrease the rank until the best possible model is determined, as well as until it is significant.

Wind and temperature are among the climate elements that have the most influence on cotton productivity in the main production regions of Azerbaijan (Figure 3). These two variables showed positive correlation and high significance in the period from sowing to flowering. Excessive rainfall and excess water during the flowering period can lead to a decrease in cotton yield. Water shortage and maximum temperature above 34°C results in a sharp decrease in cotton productivity. Using machine algorithms, it is possible to accurately predict cotton yields in major producing regions. The best algorithm was Polynomial and the least performing algorithm was Random Forest. The polynomial algorithm is successful in predicting cotton yield with climate data from planting to flowering. It is possible to predict about 90-100 days, which gives the manufacturer enough time to plan the product.

5. Conclusion

The findings of this study demonstrate that considerable changes in both climatic variables and cotton indices occur over a 5-year period. According to the diverse behaviors of climate variables at various locations, the effect of their shifting patterns on cotton output is evaluated. We examined the data on cotton yield and daily climatic factors from May to September during vegetation process. Analysis of the effects of climate variables on cotton yield was done using the regression models. The R2 value, which is within the range of the regression equations for yield, was used to pick the regression equations.

We intend to gather more information (climate, soil) in the future to increase the model's accuracy. It will be more effective in the future with the study of other weather parameters and the addition of soil types.

References

Bange, M., Baker, J. T., Bauer, P. J., Broughton, K. J., Constable, G. A., Luo, Q., Oosterhuis, D. M., Osanai, Y., Payton, P., Tissue, K. R., and Singh, B. K. (2016). Climate Change and Cotton Production in Modern Farming Systems, CAB International, Boston, MA, 61

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

Chen, T., & Guestrin, C. (2016, August). Xgboost: A scalable tree boosting system. In Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining (pp. 785-794).

Crane-Droesch, A. (2018). Machine learning methods for crop yield prediction and climate change impact assessment in agriculture. Environmental Research Letters, 13(11), 114003.

Gonzalez-Sanchez, A., Frausto-Solis, J., & Ojeda-Bustamante, W. (2014). Attribute selection impact on linear and nonlinear regression models for crop yield prediction. The Scientific World Journal, 1-10. https://doi.org/10.1155/2014/509429.

https://power.larc.nasa.gov/data-access-viewer/

https://stat.gov.az/source/agriculture/

Jans, Y., von Bloh, W., Schaphoff, S., & Müller, C. (2021). Global cotton production under climate change–Implications for yield and water consumption. Hydrology and Earth System Sciences, 25(4), 2027-2044.

Khan, P. W., Byun, Y. C., Lee, S. J., Kang, D. H., Kang, J. Y., & Park, H. S. (2020). Machine learning-based approach to predict energy consumption of renewable and nonrenewable power sources. Energies, 13(18), 4870.

Mohammad, M. Y., & Ahmed, A. D. (2022). Estimating Parameters via L-Linear Method for Second-Order Regression of Polynomial Model. Journal of Economics and Administrative Sciences, 28(134), 160-167. https://doi.org/10.33095/jeas.v28i134.2428.

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

Pambiqçiliq. (2017). Ministry of Education of the Republic of Azerbaijan, Baku.

Seyidaliyev, N. (2012). Pambıqçılığın əsasları, Baku, East-West

Shahhosseini, M., Hu, G., Huber, I., & Archontoulis, S. V. (2021). Coupling machine learning and crop modeling improves crop yield prediction in the US Corn Belt. Scientific reports, 11(1), 1606.

Shaik, M. A., Manoharan, G., Prashanth, B., Akhil, N., Akash, A., & Reddy, T. R. S. (2022, May). Prediction of crop yield using machine learning. In AIP Conference Proceedings (Vol. 2418, No. 1). AIP Publishing.

Silva, M. T., Andrade, A. S. D., Serrão, E. A. D. O., da Silva, V. D. P., & Souza, E. P. D. (2021). Applıcatıon of spatial modeling for upland cotton yield ın the semi-arid of Paraíba State, Brazil. Engenharia Agrícola, 41, 609-618.

Snider, J. L., Oosterhuis, D. M., Skulman, B. W., & Kawakami, E. M. (2009). Heat stress-induced limitations to reproductive success in Gossypium hirsutum. Physiologia plantarum, 137(2), 125-138.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., & Hofner, B. (2018). Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. Statistics and Computing, 28, 673-687. https://do.iorg/10.1007/s11222-017-9754-6.

Turner, N. C., Hearn, A. B., Begg, J. E., & Constable, G. A. (1986). Cotton (Gossypium hirsutum L.): Physiological and morphological responses to water deficits and their relationship to yield. Field Crops Research, 14, 153-170.