Journal of Agricultural Sciences and Sustainable Development

Open Access Journal
https://jassd.journals.ekb.eg/

ISSN (Print): 3009-6375; ISSN (Online): 3009-6219

Analyzing Soil Quality and Fertility in Agriculture: A Comprehensive Review of Regression Techniques

Ali, M. Z.[1], Rizk, F. H.[2], Eid, M. M.[3], Ibrahim, A.[4], Abdelhamid, A. A.[5], Khafaga, D. S.[6], Alhussan, A. A.[6], Mashaal, A. A.[7] and EL-Kenawy, E. M.[1*]

1- Department of Communications and Electronics, Delta Higher Institute of Engineering and Technology (DHIET), Mansoura 35111, Egypt

2- Computer Science and Intelligent Systems Research Center, Blacksburg 24060, Virginia, USA

3- Faculty of Artificial Intelligence, Delta University for Science and Technology, Mansoura 35111, Egypt

4- School of ICT, Faculty of Engineering, Design and Information & Communications Technology (EDICT), Bahrain Polytechnic, PO Box 33349, Isa Town, Bahrain

5- Department of Computer Science, Faculty of Computer and Information Sciences, Ain Shams University, 11566, Cairo, Egypt

6- Department of Computer Sciences, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, P.O. Box 84428, Riyadh 11671, Saudi Arabia

7- Department of Financial and Accounting Management Programs, Applied College, Princess Nora bint Abdul Rahman University, Saudi Arabia

Abstract

Agricultural systems are very complicated mechanisms of connections between plants, animals, and the biochemical processes in a way that they provide the main source of crop production, ecosystem stability, and environmental sustainability. Soil is the foundation on which farming rests, with plants growing efficiently and ecosystems functioning. Therefore, there is a need for the assessment of soil quality so as to ensure that agriculture is fruitful, stable, and sustainable. Soil gastric enzymes operate as key catalysts in chemical reactions and nutrient cycling, organic matter decay process, and soil fertility. This survey discusses the two main enzymes, amylase and urease, that play an essential role in nutrient absorption by breaking down starch and improving the nitrogen cycle. Soil physicochemical properties, land use, and weather conditions provide stability of the enzyme activity. Regression analysis techniques like multiple linear regression (MLR) and random forest (RF) machine learning classifiers use large amounts of data to explore enzymatic activity in the soil and investigate its associations with soil properties and management practices. Regression analysis additionally descends over soil enzymology to crop yield forecasting, greenhouse gas emissions accounting, and environmental degradation evaluation, among other things.

Manuscript Information:

*Corresponding authors: **EL-Kenawy, E. M.**

E-mail: skenawy@ieee.org

Received: 15/04/2024

Revised: 30/05/2024

Accepted: 02/07/2024

Published: 11/07/2024

DOI: 10.21608/JASSD.2024.283055.1018

Keywords: Agriculture; Regression Analysis; Soil Quality; Forecasting; Machine Learning.

INTRODUCTION:

Agricultural ecosystems make up a vast network of inactive interactions among elements, including plants, organisms and biochemical and biophysical processes, which tend to utilize all these activities in the growth of crops, the maintenance of ecological function and the protection of environmental integrity. Soil is a significant part of an agricultural system [1,2], which is the supporting agent of plant growth as well as comprehensive or ecological system functioning [3]. The soil quality [4], which means a variety of attitudes from the physicochemical process to biological, is contained in the assessment of agricultural production [5], resilience, and sustainability. Biological gastric enzymes occupied on the soil are the key to sustainable soil health[6]; they operate as the biological catalysts that direct the unpretentious chemical reaction, the nutrient cycle[7], organic matter turnover[8], and the fertility of the soil[9]. Amylase and urease are enzymes that carry out two important functions necessary for the nutritious diet of plants and the functioning of the soil ecosystem. They are among the many enzymes the soil contains [10], which makes them a major one. An enzyme called amylase, in a way, accelerates the depolymerization of starch into simpler sugars that aid the uptake of nutrients for both plants and microbial communities [11]. However, urease catalyzes the conversion of urea into ammonium and carbon dioxide, thereby contributing to the cycling of nitrogen, which in turn enhances soil fertility.

Soil physicochemical properties [12], land management zoning, and given environment all determine the activity of these enzymes. The improvement of the soil's physico-chemical features is one of the most imperative steps for bringing about the best enzymatic activity and general healthiness of the soil. Major types of parameters are soil pH, electrical conductivity, bulk density, organic matter content, and nutrient availability [13,14], which play key roles in enzyme activities. Such parameters dictate the chemical structure of soils used for farming, and they determine how well soils support agriculture and other environmental processes. Successful determination of the intricate relationships between soil content biophysical attributes and enzyme activities requires a complete understanding in order to guide the establishment of sustainable land utilization methods and enhance crop productivity. The regression analysis models give us valuable tools that assist in exposing the inner intricacy of the interrelations among soil features[15], land management methods, and enzyme dynamics. In recent years, there has been an increasing popularity of multi-linear regression (MLR) and more sophisticated machine learning methods[16], including random forest (RF), which are applied to the analysis of large data sets of multiple samples and the investigation of electronic soil enzymatic activity.

What has already been mentioned helps us see the delicate and nuanced links between the fertility of the soil and the activity of enzymes. These tools yield valuable information regarding how enzymatic behavior in soil functions, thereby allowing researchers to identify important precursors and key factors for enzyme dynamics. By combining field research and theory and incorporating viewpoints from other disciplines, this review will aim to provide a full map of the

complex interplay of the physicochemical characteristics of soil and enzyme activity in agriculture. The recognition of the roles of factors that affect and the implications of soil enzyme dynamics in the formation of the basis and sustainability of agriculture can be completed by investigating the latest research findings and methods employed in the evaluation of amylase and urease activities. In addition to this, the study learns the variety of fields that regression model analysis can permeate in agricultural research beyond soil enzymology[17]. Regression analysis has proven its ability to be operated in a fair number of fields, such as determining crop yields, monitoring greenhouse gas emissions, and figuring out the cause of heavy metal pollution in agricultural soils[18]. It can help explain complex environmental processes and provide research-based propositions, which in turn can facilitate sustainable land management methods[19]. The whole globe is faced with deadly climate changes, population increases, and environmental declines and sustainable agriculture presents an exceptionally pressing need for solutions. Through the use of multivariate statistical models and considering disciplinary approaches[20], scientists and professionals can make novel insights on soil enzymology and agricultural produce and the resilience of an ecosystem.

This study will achieve its objectives by applying empirical facts and theoretical frameworks, which are an interdisciplinary approach. This study will thus contribute to the scholarly field of agricultural research with the aim of facilitating the transition toward a sustainable future for both agriculture and society at large.

LITERATURE REVIEW:

The restoration of the physicochemical parameters and soil structure are crucial as they determine how extracellular enzymes perform a wide range of biochemical processes that are vital for the whole ecosystem. This study is to evaluate the soil enzyme amylase and urease activities with the use of both multiple linear regressions (MLR) and random forest (RF). It is a combination of components, including pH, soil electrical conductivity, bulk density, total nitrogen, total phosphorus, soil organic carbon, and soil water content. The urease levels are higher in disturbed land, rape land, and fishponds than in other surrounding areas.

Amylase, which is an enzyme of bacterial type, is an essential microorganism, giving high productivity in fish ponds. Nitrogen total and soil water level are considered to be indicator features by the machine learning model, which influence enzyme activity significantly following the comparison with those attributes that are less likely to influence enzymes. Interestingly, the random forest also gives confidence with regard to extrapolation of soil enzyme activity during land-use changes, pinpointing the nonlinear impacts better, this being stronger than MLR.

SUMMARY TABLE:

The table below reveals that a lot of studies have been done on using statistics and machine learning in agriculture. It looks at a number of things, from the type of soil (like enzyme activity and fertility) to the impact of farming methods in West Africa on CO2 levels. On the other hand, some studies watch the possibility of using photos from satellites to better control crops and whether the prices of agriculture futures and crude oil move together (whether they move in the same direction in lightning). It is not only in terms of

efficacy but also in health-based farming that phytoremediation is an exciting possibility. The results of the studies presented prove how fruitful the field of agricultural information is for increased output, environmental protection, and sustainable farming practices.

Table 1: Summary of Related Works.

Ref.	Focus	Methodology	Key Findings
[21]	Influence of land use on soil enzyme activities	MLR and RF models	Bean land, rappel land and fish pond enzyme assays reveal higher urease and amylase bacterial enzyme activity; these enzymes are conversed into environmentally toxic products. Enzymatic activity is principally regionalized in terms of the degree of water saturation and total nitrogen. The RF model goes on to appear above the MLR model in the matter of handling nonlinear phenomena quite clearly.
[22]	Factors affecting CO2 emissions in West Africa	Panel quantile regression with non-additive fixed effects	The practice of traditional farming kindles a drop-off in the emission of CO2 from the liquids. In spite of the fact that intensive rain and crop production intensified the total emission rate. Divergence in CO2 emissions in West Africa depends on the economic level of the people living there.
[23]	Quantification of crop traits using EnMAP satellite data	Hybrid Retrieval Approach with machine learning algorithms	I modeled with high accuracy for biochemical and biophysical characteristics. Of various approaches, ANN weighs in on top for all metrics, namely the accuracy, model size and execution speed. Capable of producing Cab and LAI predictions.
[24]	Source identification of heavy metal(loid)s in agricultural soils	PMF model, regression modeling, geospatial mapping	Soil contamination with heavy metals(loids) is mostly due to lithogenic and anthropogenic sources. Although the PMF model is helpful in identifying the sources, for an improvement in precision, other models have to be integrated.
[25]	Prediction of soil GHG emissions using ML models	ML regression models (graphical, shallow, deep learning)	The LSTM model seems to be quite competent as far as CO2 and N2O emissions estimation is concerned, which is kind of an indication of the potential of ML algorithms to improve environmental science research.

[26]	Estimation of crop canopy cover from NDVI data	Meta-analysis	Even though CC may be assessed with a small quantity of accuracy based on the data of NDVI, it is still necessary to confirm it locally. A simple equation between the proposed drought index and the calibration of the AquaCrop model converts NDVI into a good parameter.
[27]	Soil fertility evaluation for precision agriculture	Partial least squares regression	High accuracy analysis is needed to find the features of the soil that are vital for agricultural purposes. The audience will get a glimpse of ways in which past farming data and machine learning can invent new kinds of agriculture methods.
[28]	Groundwater suitability assessment for irrigation	PLSR model and GIS software	Due to the integration of physicochemical characteristics, water quality markers, PLSR model, and GIS, understanding the water quality status for irrigation turned out to be an easier task.
[29]	Impact of crude oil price on agricultural futures in China	Quantile-on-quantile statistics, stochastic volatility model	Crude oil and farm futures are not equally dependent on volatility. This shows how important it is to make smart decisions in policy and finance.
[30]	Metal absorption efficiency in Forage Sorghum	Prediction systems	The roots of a plant take in more metals than any other part. For some metals, BAF readings are higher than 1, which means that phytoremediation might be possible.

REVIEW:

Restoration and reclamation of the coastal tidal land influence the soil structure through which physicochemical conditions vary tremendously. In addition, they affect soil extracellular enzymes, the enzymes needed for the soil ecosystem to function and take part in a lot of biochemical processes. For an estimation of soil amylase and urease activities, [21] utilizes multiple linear regressions (MLR) and random forest (RF) models with such parameters as pH, electrical conductivity, bulk density, total nitrogen, total phosphorus, soil organic carbon and soil water content. Releases indicate that bacterial enzyme activity of amylase is many times higher in fishponds. In contrast, urease activity turns out to be especially high in rape land, broad bean land, and fishponds. The total nitrogen content in soils and the soil water content emerge as the key determinants of urease and amylase activity according to the random forest model. Importantly, the R^2 value and the error index grow smaller, and so does the random forest model's performance compare to the MLR model. The model development stands out in excellence in handling nonlinear effects and protecting against noise and overfit. It is a robust technique to extrapolate soil enzyme activity against land use changes. The leading industry in West African economies, which is cultivation, makes up a

substantial proportion of the region's greenhouse gas emissions mission accomplished. [22] highlights the major factors that lead West Africa to have low CO2 emissions, intermediate CO2 emissions or high CO2 emissions in order to realize the level of environmental degradation. Input**: It contains these important items, like industrialization, agriculture, the use of renewable energy, and economic growth. The piece incorporates quintile decomposition analyzing panel data for fifteen ECOWAS countries from 90 to 2015 by using panel quantile regression with non-additive fixed effects. On the one hand, the outcomes demonstrated that agricultural production results in significant emissions reduction from liquid sources. On the other hand, it turned out that the production implies an overall rise in emissions, signaling the trend of the transformation of farming practices into the direction of classical methods and biomass utilization from agricultural waste for energy production. The analysis demonstrates the diverse nature of these impacts by pointing out significant variations in the environmental conditioning factors in different CO2 emitter categories. As a result, the quantile decomposition data explains the detailed CO2 emissions by mainly segregating income groups in Western Africa, including a higher emission level between lower-middle income and lower-income economies at higher quantiles. Sophisticated processing tools are required to separate valuable information at the image spectroscopical level, particularly in simultaneous operations nowadays. It researches the possible utilization of an advanced scientific onboard processor accompanying the next-generation EnMAP satellite in the quantification of spectroradiometric images, which are bio-physical and bio-chemical. Quick and effective pre-trained learning models based on the lookup database of spectra of synthetic vegetation and the radiative transfer model, PROSAIL, are applied together using the Hybrid Retrieval Approach. The widely-used model offers space information on key crop traits such as Canopy Chlorophyll Content, Canopy Leaf Area Index, and Crops' Leaf Angle of Inclination. In this way, the approach is relevant to any site without having to collect data specific to that site. High-performance estimations for biophysical and biochemical variables were shown by four machine learning algorithms: artificial neural networks (ANN), random forest regression (RFR), support vector machine (SVM), and Gaussian process regression (GPR). When it comes to ANNs, the accuracy, model size, and execution time were the highest among other types of algorithms. Functional characterization, including LAI and Cab, to a great extent, was shown to be strong enough, as validated with the SPARC03 dataset of Barrax. The ANN model showed an RMSE of 0.81 m^2 m^{-2} and 6.2 μg cm^{-2}, respectively. After the simultaneous expression of crop traits in EnMAP-simulating conditions, realistic within-field geometry could be observed. Compliance of the LAI result estimated in the SNAP biophysical process with the required accuracy confirms the applicability of ANN models for developing hybrid crop trait detection systems in the future based on the satellite imaging spectroscopy data faithful for practical crop monitoring missions, especially for grassland and vegetative phases of maize[23]. Figuring out where the origin of heavy metals (loids), natural elements of the Earth crust find their way to the soils, is not an easy task either. Because the positive matrix factorization

(PMF) model in combination with regression modeling and geospatial mapping is applied for the purpose of finding heavy metal(loid) sources in agricultural soils of Handan which is larger than 12,000 km^{2}, [24]aims to find out if PMF model is applicable for such large area. Moreover, surface soils had a significant increase of Cd, Cu, Pb and Zn; of these, the highest proportion of the total potential risks was accounted by Cd at 73%. The PMF model showed that lithogenic sources were the main source of Fe (71.8%), Cr (60.0%), V (52.9%), Cu (50.7%), Ni (42.2%) and Mn (41.4%) whereas industrial sources were the major lead (47.8%) and Cd (56.9%) sources in agricultural so The results of this investigation established a combined input of Co (54.1%), As (42.9%), and Zn (40.0%) from natural background, agriculture, and vehicle emissions. Uncertainty analysis revealed to be decisive as the levels of heavy metal(loids) contributed by different pollution sources strongly differed, according to the PMF model. Although PMF helped ascribe qualitative sources to their spatial distributions, the authors call upon its integration with reacting transport models and emission databases are crucial to getting more precise and conclusive results on the provenance of heavy metal(loids) in the soil. Big temporal and spatial issues of complex events render machine learning (ML) models quite attractive of late. [25] looks at the prediction of soil greenhouse gas (GHG) emissions from an agricultural field in Quebec, Canada, using three different types of machine learning (ML) regression models: graphical or statistical regression, shallow learning, and deep learning. The location is a nitrous oxide (N2O) and carbon dioxide (CO2) flux recorder by collecting data over a course of five years with the addition of a multitude of environmental, agronomic, and soil data. The outcome of the experiment which compared flow estimates and cross-validation with statistical data revealed that LSTM (Long Short Term Memory) model performed the best for machine learning application. Being the model which had the greatest correlation coefficient and which had the lowest root mean square error value for both CO2 ($R = 0.87$, RMSE = 30.3 mg ·m−2·hr−1) and N2O flow predictions ($R = 0.86$, RMSE = 0.19 mg ·m−2·hr−1), LSTM became the one Even more unexpectedly, the Root Mean Square (R2) from the LSTM, which was a neural-network based model, was more effective than RZWQM2 which was a biophysics-based model employed in a previous study. While the proposed classical regression models (RF, SVM and LASSO) might efficient to predict fluctuations in CO2 fluxes that are cyclical and seasonal, the forecast of the N2O peak values prove to be more difficult. The overparameterization issue could be improved with optimized hyperparameter search and so was the error susceptibility in forecasting GHG emissions. Through such exhaustive study, the LSTM model becomes clear for simulating greenhouse gas emissions from soils of agricultural fields. This comparison also opens the door to the possible use of ML algorithms in predicting GHG release into the atmosphere for environmental scientists. Vegetation cover calculated in terms of crop canopy cover (CC) is a critical factor upon which crop development and model calibrations depend. However, the CC can be derived from normalized difference vegetation index (NDVI) data collected by satellites. [26] elucidates a puzzling but must-solve matter of calculating CC. By employing 19 studies

subjecting to 1397 observations, a complete meta-analysis is conducted and proposes generic models for 13 specified crops. The research discloses uncertainties with their range from 6% to 18% and indicates the possible incorrect estimation of actual CC. Very often, the r2 values between 0.75 and 1 are interpreted as the relationships being widely acknowledged and considered satisfactory. Coefficient estimates not only assure us of the presence of properly estimated relationships but also make the equation multilinear, thereby canceling out underfitting or overfitting. Potential non-sampling mistakes must be considered when overlooking the situations that were not investigated in the research. This note on validation at the local level also accentuates the reasonableness of taking into account the natural differences that exist across the different parts of the research area. The suitability of NDVI as an estimate of drought index and calibration of the AquaCrop model is demonstrated, while the case study results for wheat are referred to as an example. The simulation outcome employed reference data from three cultivating seasons to validate the model functionality, which indicated good potential for utilizing NDVI-CC for research. On the other hand, local validation remains necessary for extrapolation. The overall potential for modeling applications is shown by acceptable Root Mean Squared Errors (RMSE) for sensitive analysis that equals the predetermined levels. With the development of precise farming technologies, nowadays, the level of adoption of precision agriculture is getting higher and higher. It applies fascinating sensors to determine the nutrient level of the soil as well as to define the requirements of the plants. Machine learning algorithms are employed to shape the framework of prediction using the model. Sustainable measures aim to achieve the highest possible productivity of agriculture while minimizing the environmental footprint. In [27], a partial least squares approach is proposed in order to lead the soil fertility evaluation that takes context into account.Moreover, the analysis offers yield data for a specific location by using the climate events that occurred between 2001 and 2015. The analysis reveals the Pearson correlation coefficient (R2) of 0.9189, the mean square error of cross-validation of 0.54 T/ha, and the mean square error of calibration (RMSEC) of 0.20 T/ha. Of particular importance is the model's result, which displays high accuracy with the R2 0.9345 and RMSECV 0.54% for the prediction of organic matter and R2 0.9239 and RMSECV 5.28% for the prediction of clay. This complete method exhibits the possibility to use historical context and new modeling strategies to correctly analyze and predict baseline soil properties that are of much importance to boosting agricultural capability. The excellent performance of the proposed model conveys how valuable it is for the purpose of making farming methods that are both economical and environmentally friendly, thus keeping in line with the main theme of precision agriculture. The trial employed a group of IWQIs, (IWQIs), (TDS, SAR, PS, MH, and RSC) as indicators for assessing the agricultural suitability of the alluvial aquifer of Makkah Al-Mukarramah Province, Saudi Arabia. Different cation and anion contents, Ca–HCO3, Na–Cl, and Ca–Mg–Cl–SO4 facies, inferred to be a result of evaporation, saltwater intrusion, and the reverse ion-exchange phenomena, were pointed out in 114 groundwater wells after detailed investigation.

The computerized IWQI model managed how I should choose the possible crops by labeling the soil samples into nil, low to moderate, and high to severe classes of irrigation limitations. Moreover, a partial least squares regression model with strong R2 values varying from 0.72 to 1.00 in the validation datasets proved the estimation of the six classes of water quality with success. The research design is based on the objective of completely evaluating groundwater suitability for irrigation and understanding the factors influencing the chemical-water quality in the arid-semiarid regions by combining the application of physicochemical properties, water quality indices, PLSR model and GIS software[28]. The present work explores the response of rice and soybeans futures in China to the crude oil price, which is measured by quantile-on-quantile statistics. The paper considers the stochastic volatility model as the mean of conditional variance, but the time-varying parameters are employed to assess the potentially fluctuating volatility and the varied types of dependency among quantiles. The data provide the growing trend toward the absolute volatility spillover of the agricultural volatility with the quantiles (numbers of quotations or ranges of figures) greater and greater variability degrees of dependency between the crude oil volatility from and storm in the Chinese agricultural futures. Under the financial market's sudden changes as well as stable market conditions, it can be seen that the volatility dependency is asymmetric. Surprisingly, it is significant that during a regular mode of crude oil market, no apparent effect of oil volatility is seen. So, there are very excessive or very small quantiles of oil volatility, which are the leading factors of agricultural volatility. The analysis as well discovers that the volatility exchange's behavior is deeply connected. The influence of volatility on returns also shows clear time variation. This draws a sign that tactful decision-making facing the change in market situation has broad economic implications for portfolio managers and policymakers who act in various fields of finance and authority in the world[29]. [30] modeled by utilizing prediction systems to evaluate the efficiency of the leaves, stems and roots of Forage Sorghum in absorbing nine metals (Cd, Co, Cr, Cu, Fe, Mn, Ni, Pb, and Zn) at different soil mixtures obtained by blending it with poultry manure (0, 10, 20, 30, and The researchers' results showed that the roots had more metals than others parts of the plant. Also, bioaccumulation factor (BAF) values were computed; BAF values for Cr, Fe, Ni, Pb, and Zn were less than 1, whereas BAF values for Co, Cu, Mn, and Cd were higher than 1. However, Co, Cr, and Ni were the only metals that outperformed the value of the translocation factor by less than one, which implies limited translocation to the leaves (or stems). Our results showed that the pH of the soil was negatively correlated with the metals extractable from plant parts but positively linked to EC and the amount of organic matter in the sample. The analysis showed that the predicted and experimentally observed metal contents of the three plant tissue components varied at the same level of precision, which denotes the high predictability of the models. Therefore, this modeling will offer more credence in estimating the potential risks of feeding sorghum hay supplemented with poultry fertilizer among human health effects. The study concludes by focusing on the significance of the land use changes for the activity of soil enzymes in the

tidal coastal zone. The random forest model efficiently predicts enzyme activity, and it is the best technique for handling complicated interactions. The understanding of these biochemical reactions is paramount for the assessment of ecosystem health and designing green technologies as shoreline management tools. The outcome adds information about the intricate mechanisms of enzyme enclosure in soils. Such knowledge will enable the stakeholders, like the land managers, environmental scientists and legislators, to make the right decisions concerning preservation and restoration of the coastal ecosystems.

REGRESSION ANALYSIS MODELS FOR: CROP YIELD PREDICTION:

Regression analysis models become ordinary tools used in agriculture as strong methods of crop yield forecasting[31], often based on multiple major parameters. The main task of the models is to find a linear relationship between an independent variable or group of variables, such as weather patterns, soil properties[32], and the farming methods employed, and the dependent variable, that is, crop yield.

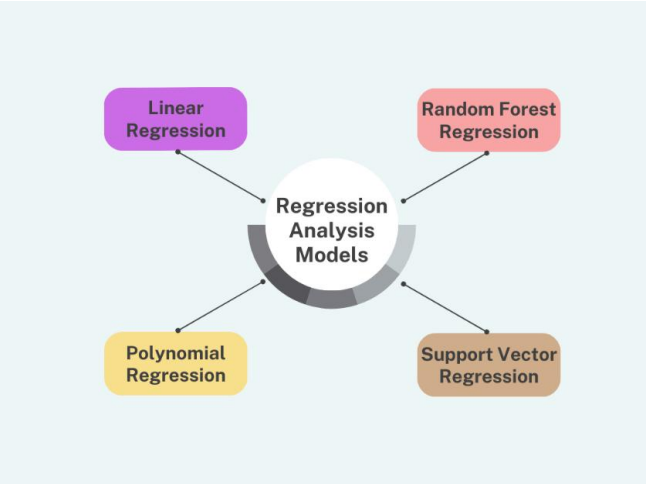

Figure 1: Overview of Regression Analysis Models

Linear Regression: Linear regression[33], a widely adopted simple method of predicting agricultural productivity, is used by many farmers. The hypothesis is that it is assumed that a relationship exists on the same line, connecting the independent factors to crop production. In this model, the expected yield is computed by adding the variables, which all have their multiplication by the appropriate coefficients. Linear regression is a simple statistical technique that can help to understand the direction of variation and the size of an association between variables. Nevertheless, we need to understand that it might stylize complex relationships.

Polynomial Regression: Polynomial regression is an improvement of linear regression by the development of relationships, particularly between the independent and dependent variables. The regression equation includes polynomial terms, including squared and cubed terms, to interact with the data and fit to a curve. The capability of a polynomial regression model in dealing with complicated correlations hidden in agricultural data is due to the model`s adaptability. Holistically, using the higher-order polynomials may bring about model complexity and overfitting [34,35].

Random Forest Regression: The random forest regression technique [36], which is a machine learning technique based on an ensemble of decision trees, is being exploited to predict agricultural yields. Developing forests undergo what is referred to as the tree training process, which involves specific segments of the data being used to train individual trees within the forest. The final prediction is then made by combining all three predictions and producing a final output. The random forest regression [37], it is worth mentioning, is capable of dealing with non-linear correlation and a reasonable amount of

interaction among variables and, hence, is very efficient and indeed an ideal tool to use for complex agricultural datasets. Furthermore, it also lacks the overfitting problem and has the ability to handle abnormal values and drop-outs that are not useful.

Support Vector Regression: Support vector regression (SVR) is a machine learning method applied for the forecasting of crop production. Support Vector Regression (SVR) also employs innovative yet intuitive vector transformation techniques that enable the mapping of input variables into a space with higher dimensions. The next step of the process is to construct the hyperplane, which is closer to the data and omits errors across the process. Support vector machines (SVR) are very effective when application data is high-dimensional and contains nonlinear relationships. The parameter setting of Support Vector Regression (SVR) may be challenging as the duration or type of the kernel function can play a role [38,39].

DATA SOURCES AND CHALLENGES:

Variables, specifically histrionic yield data, weather data, soil details and management practices, are being incorporated into the regression analysis models to foretell the crop yield. Two main difficulties linked to the construction of these models are scarcity of data actuators, concerns regarding data quality [40], the existence of multicollinearity among predictors, and up-to-date model updates to adapt to transforming environmental issues and the efficaciousness of management techniques. Eventually, regression analysis models will be outstanding for forecasting crop yields and providing relevant information for policy-related actions in the field of agriculture. Conditioning the suggestion of a model on the unique features of the dataset and the specific purpose of the research will enable one decision to be made over another, as the model of each type has its strong and weak points. In this context, researchers and practitioners can add considerable impetus to the accuracy of crop yield forecasting and thus shape the sustainability of the agricultural sector by executing the models properly and addressing the problems that may arise.

RESULTS AND DISCUSSION:

The soil's amylase and urease activities were analyzed together with physicochemical attributes, which demonstrated potential correlations. It was that soil pH, electrical conductivity, bulk density, and organic matter content were the main variables affecting enzyme activities in both MLR and RF models. pH levels proved to be highly relevant with respect to both salivary amylase and urease activities, as well as the effects of soil salinity, which was represented by electrical conductivity, on urease activity. On the other side, the enzyme activities correlate with the bulk density negatively, indicating that in the cases of compacted soils, it is difficult for microorganisms to proliferate. It was noticed that the higher organic matter content resulted in higher enzyme activities, and this condition emphasizes the fact that organic matter is beneficial for maintaining healthy soils. It was not unexpected that enzyme activities responded differently to land use types, and fishponds demonstrated an enhanced amylase activity, whereas other areas, such as land disturbances, rape plantation and fishponds, had higher urease activity levels. RF model had a better success rate in predicting enzyme activities due to its ability to handle nonlinear effects and extrapolation of the

changes during land use patterns. Soil nitrogen total content and water saturation achieved statistical significance as main contributors to enzyme activities. Both regression analysis and models were later used in agricultural research, among other things, in crop harvest prediction, soil fertility determination, and environmental impact assessment, where they exhibited versatility and effectiveness in addressing relatively complicated agricultural issues.

CONCLUSION:

In general, this review demonstrates the applicability of regression analysis models in different ways in agricultural studies. Following an examination of several research studies, it becomes obvious that regression analysis is a flexible tool that can be adopted to find solutions to complex agricultural problems .A combination of familiar conventional regression techniques with advanced machine learning algorithms, like random forest regression and support vector regression, promotes agriculture research through innovation. The models can help in better understanding of the complex scenarios, detect hidden and non-linear relationships as well as provide accurate forecasts of detailed environmental causes. Collecting information from various sources for future development and trying regression analysis models into farming decision-making processes may be proven to help increase production, reduce environmental hazards, and choose sustainable farming methods. Applying the regression analysis approach provided an opportunity for scholars and professionals to review agriculture production at different levels in terms of its complexity, hence boosting the performance of the whole food systems around the world. Thus, regression models have proved to be effective, and they are seen as a key solution for the challenges facing the agricultural sector of the modern world. The models that keep evolving with the help of studies and practical implementation are able to really change agricultural methodologies significantly. This is regarded as one of the most important features of these models – the next step is to create a resilient and sustainable future for both the agricultural sector and society at large.

REFERENCES:

[1] Sharma, R., Parhi, S., & Shishodia, A. (2021). Industry 4.0 Applications in Agriculture: Cyber-Physical Agricultural Systems (CPASs). In V. R. Kalamkar & K. Monkova (Eds.), Advances in Mechanical Engineering (pp. 807–813). Springer. https://doi.org/10.1007/978-981-15-3639-7_97

[2] Khatoon, Z., Huang, S., Rafique, M., Fakhar, A., Kamran, M. A., & Santoyo, G. (2020). Unlocking the potential of plant growth-promoting rhizobacteria on soil health and the sustainability of agricultural systems. Journal of Environmental Management, 273, 111118. https://doi.org/10.1016/j.jenvman.2020.111118

[3] Applied Sciences | Free Full-Text | Can We Use Functional Annotation of Prokaryotic Taxa (FAPROTAX) to Assign the Ecological Functions of Soil Bacteria? (n.d.). Retrieved April 1, 2024, from https://www.mdpi.com/2076-3417/11/2/688

[4] Wubie, M. A., & Assen, M. (2020). Effects of land cover changes and slope gradient on soil quality in the Gumara watershed, Lake Tana basin of North–West Ethiopia. Modeling Earth Systems and Environment, 6(1), 85–97. https://doi.org/10.1007/s40808-019-00660-5

[5] Jung, J., Maeda, M., Chang, A., Bhandari, M., Ashapure, A., & Landivar-Bowles, J. (2021). The potential of remote sensing and artificial intelligence as tools to improve the resilience of agriculture production systems. Current Opinion in Biotechnology, 70, 15–22. https://doi.org/10.1016/j.copbio.2020.09.003

[6] Javed, Z., Tripathi, G. D., Mishra, M., & Dashora, K. (2021). Actinomycetes – The microbial machinery for the organic-cycling, plant growth, and sustainable soil health. Biocatalysis and Agricultural Biotechnology, 31, 101893. https://doi.org/10.1016/j.bcab.2020.101893

[7] Dai, Z., Xiong, X., Zhu, H., Xu, H., Leng, P., Li, J., Tang, C., & Xu, J. (2021). Association of biochar properties with changes in soil bacterial, fungal and fauna communities and nutrient cycling processes. Biochar, 3(3), 239–254. https://doi.org/10.1007/s42773-021-00099-x

[8] Marschner, P. (2021). Processes in submerged soils – linking redox potential, soil organic matter turnover and plants to nutrient cycling. Plant and Soil, 464(1), 1–12. https://doi.org/10.1007/s11104-021-05040-6

[9] Boudjabi, S., & Chenchouni, H. (2022). Soil fertility indicators and soil stoichiometry in semi-arid steppe rangelands. CATENA, 210, 105910. https://doi.org/10.1016/j.catena.2021.105910

[10] Lee, S.-H., Kim, M.-S., Kim, J.-G., & Kim, S.-O. (2020). Use of Soil Enzymes as Indicators for Contaminated Soil Monitoring and Sustainable Management. Sustainability, 12(19), Article 19. https://doi.org/10.3390/su12198209

[11] Bogati, K., & Walczak, M. (2022). The Impact of Drought Stress on Soil Microbial Community, Enzyme Activities and Plants. Agronomy, 12(1), Article 1. https://doi.org/10.3390/agronomy12010189

[12] Land | Free Full-Text | Effects of Soil Bund and Stone-Faced Soil Bund on Soil Physicochemical Properties and Crop Yield Under Rain-Fed Conditions of Northwest Ethiopia. (n.d.). Retrieved April 1, 2024, from https://www.mdpi.com/2073-445X/9/1/13

[13] Land Degradation & Development | Environmental & Soil Science Journal | Wiley Online Journal. (n.d.). Retrieved April 1, 2024, from https://onlinelibrary.wiley.com/doi/full/10.1002/ldr.3657

[14] Li, Y., Li, Z., Cui, S., & Zhang, Q. (2020). Trade-off between soil pH, bulk density and other soil physical properties under global no-tillage agriculture. Geoderma, 361, 114099. https://doi.org/10.1016/j.geoderma.2019.114099

[15] Kireev, T., Kukartsev, V., Pilipenko, A., Rukosueva, A., & Suetin, V. (2022). Analysis of the Influence of Factors on Flight Delays in the United States Using the Construction of a Mathematical Model and Regression Analysis. 2022 IEEE International IOT, Electronics and Mechatronics Conference (IEMTRONICS), 1–5. https://doi.org/10.1109/IEMTRONICS55184.2022.9795721

[16] Adeniji, S. E., Uba, S., & Uzairu, A. (2020). Multi-linear regression model, molecular binding interactions and ligand-based design of some prominent compounds

against Mycobacterium tuberculosis. Network Modeling Analysis in Health Informatics and Bioinformatics, 9(1), 8. https://doi.org/10.1007/s13721-019-0212-6

[17] Qiao, Z., Yao, X., Liu, X., Zhang, J., Du, Q., Zhang, F., Li, X., & Jiang, X. (2021). Transcriptomics and enzymology combined five gene expressions to reveal the responses of earthworms (Eisenia fetida) to the long-term exposure of cyantraniliprole in soil. Ecotoxicology and Environmental Safety, 209, 111824. https://doi.org/10.1016/j.ecoenv.2020.111824

[18] Fei, X., Lou, Z., Xiao, R., Ren, Z., & Lv, X. (2022). Source analysis and source-oriented risk assessment of heavy metal pollution in agricultural soils of different cultivated land qualities. Journal of Cleaner Production, 341, 130942. https://doi.org/10.1016/j.jclepro.2022.130942

[19] Yang, J., Sun, Y., Wang, Z., Gong, J., Gao, J., Tang, S., Ma, S., & Duan, Z. (2022). Heavy metal pollution in agricultural soils of a typical volcanic area: Risk assessment and source appointment. Chemosphere, 304, 135340. https://doi.org/10.1016/j.chemosphere.2022.135340

[20] Liu, L. (2022). Quantitative Impact Analysis of Financial Support on Regional Science and Technology Innovation and Productivity Based on the Multivariate Statistical Model. Mathematical Problems in Engineering, 2022, e7175807. https://doi.org/10.1155/2022/7175807

[21] Xie, X., Wu, T., Zhu, M., Jiang, G., Xu, Y., Wang, X., & Pu, L. (2021). Comparison of random forest and multiple linear regression models for estimation of soil extracellular enzyme activities in agricultural reclaimed coastal saline land. Ecological Indicators, 120, 106925. https://doi.org/10.1016/j.ecolind.2020.106925

[22] Nwaka, I. D., Nwogu, M. U., Uma, K. E., & Ike, G. N. (2020). Agricultural production and CO2 emissions from two sources in the ECOWAS region: New insights from quantile regression and decomposition analysis. Science of The Total Environment, 748, 141329. https://doi.org/10.1016/j.scitotenv.2020.141329

[23] Danner, M., Berger, K., Wocher, M., Mauser, W., & Hank, T. (2021). Efficient RTM-based training of machine learning regression algorithms to quantify biophysical & biochemical traits of agricultural crops. ISPRS Journal of Photogrammetry and Remote Sensing, 173, 278–296. https://doi.org/10.1016/j.isprsjprs.2021.01.017

[24] Yuanan, H., He, K., Sun, Z., Chen, G., & Cheng, H. (2020). Quantitative source apportionment of heavy metal(loid)s in the agricultural soils of an industrializing region and associated model uncertainty. Journal of Hazardous Materials, 391, 122244. https://doi.org/10.1016/j.jhazmat.2020.122244

[25] Hamrani, A., Akbarzadeh, A., & Madramootoo, C. A. (2020). Machine learning for predicting greenhouse gas emissions from agricultural soils. Science of The Total Environment, 741, 140338. https://doi.org/10.1016/j.scitotenv.2020.140338

[26] Tenreiro, T. R., García-Vila, M., Gómez, J. A., Jiménez-Berni, J. A., & Fereres, E. (2021). Using NDVI for the assessment of canopy cover in agricultural crops within modelling research. Computers and Electronics

in Agriculture, 182, 106038. https://doi.org/10.1016/j.compag.2021.106038

[27] Helfer, G. A., Victória Barbosa, J. L., Santos, R. dos, & da Costa, A. B. (2020). A computational model for soil fertility prediction in ubiquitous agriculture. Computers and Electronics in Agriculture, 175, 105602. https://doi.org/10.1016/j.compag.2020.105602

[28] Masoud, M., El Osta, M., Alqarawy, A., Elsayed, S., & Gad, M. (2022). Evaluation of groundwater quality for agricultural under different conditions using water quality indices, partial least squares regression models, and GIS approaches. Applied Water Science, 12(10), 244. https://doi.org/10.1007/s13201-022-01770-9

[29] Hau, L., Zhu, H., Huang, R., & Ma, X. (2020). Heterogeneous dependence between crude oil price volatility and China's agriculture commodity futures: Evidence from quantile-on-quantile regression. Energy, 213, 118781. https://doi.org/10.1016/j.energy.2020.118781

[30] Eid, E. M., Hussain, A. A., Alamri, S. A. M., Alrumman, S. A., Shaltout, K. H., Sewelam, N., Shaltout, S. K., El-Bebany, A. F., Ahmed, M. T., Al-Bakre, D. A., Alfarhan, A. H., Picó, Y., & Barcelo, D. (2023). Prediction Models Based on Soil Characteristics for Evaluation of the Accumulation Capacity of Nine Metals by Forage Sorghum Grown in Agricultural Soils Treated with Varying Amounts of Poultry Manure. Bulletin of Environmental Contamination and Toxicology, 110(1), 40. https://doi.org/10.1007/s00128-022-03654-9

[31] Paudel, D., Boogaard, H., de Wit, A., Janssen, S., Osinga, S., Pylianidis, C., & Athanasiadis, I. N. (2021). Machine learning for large-scale crop yield forecasting. Agricultural Systems, 187, 103016. https://doi.org/10.1016/j.agsy.2020.103016

[32] Wang, F., Wang, Q., Adams, C. A., Sun, Y., & Zhang, S. (2022). Effects of microplastics on soil properties: Current knowledge and future perspectives. Journal of Hazardous Materials, 424, 127531. https://doi.org/10.1016/j.jhazmat.2021.127531

[33] Lederer, J. (2022). Linear Regression. In J. Lederer (Ed.), Fundamentals of High-Dimensional Statistics: With Exercises and R Labs (pp. 37–79). Springer International Publishing. https://doi.org/10.1007/978-3-030-73792-4_2

[34] Wang, L., Liu, J., & Qian, F. (2021). Wind speed frequency distribution modeling and wind energy resource assessment based on polynomial regression model. International Journal of Electrical Power & Energy Systems, 130, 106964. https://doi.org/10.1016/j.ijepes.2021.106964

[35] Shi, M., Hu, W., Li, M., Zhang, J., Song, X., & Sun, W. (2023). Ensemble regression based on polynomial regression-based decision tree and its application in the in-situ data of tunnel boring machine. Mechanical Systems and Signal Processing, 188, 110022. https://doi.org/10.1016/j.ymssp.2022.110022

[36] Iqbal, M., Onyelowe, K. C., & Jalal, F. E. (2021). Smart computing models of California bearing ratio, unconfined compressive strength, and resistance value of activated ash-modified soft clay soil with adaptive neuro-fuzzy inference system and ensemble random forest regression techniques. Multiscale and Multidisciplinary Modeling, Experiments and

Design, 4(3), 207–225. https://doi.org/10.1007/s41939-021-00092-8

[37] Wang, F., Wang, Y., Zhang, K., Hu, M., Weng, Q., & Zhang, H. (2021). Spatial heterogeneity modeling of water quality based on random forest regression and model interpretation. Environmental Research, 202, 111660. https://doi.org/10.1016/j.envres.2021.111660

[38] Dash, R. K., Nguyen, T. N., Cengiz, K., & Sharma, A. (2023). Fine-tuned support vector regression model for stock predictions. Neural Computing and Applications, 35(32), 23295–23309. https://doi.org/10.1007/s00521-021-05842-w

[39] Najafzadeh, M., & Niazmardi, S. (2021). A Novel Multiple-Kernel Support Vector Regression Algorithm for Estimation of Water Quality Parameters. Natural Resources Research, 30(5), 3761–3775. https://doi.org/10.1007/s11053-021-09895-5

[40] Peer, E., Rothschild, D., Gordon, A., Evernden, Z., & Damer, E. (2022). Data quality of platforms and panels for online behavioral research. Behavior Research Methods, 54(4), 1643–1662. https://doi.org/10.3758/s13428-021-01694-3