Research on soil moisture prediction based on VAR-ARIMA model

Xin Wen [1,*], Juan Wei [2], Jinyang Zhang [3], Junrong Yue [4]

[1] School of Economics and Management, Fuzhou University, Fujian, China

[2] School of Business and Tourism, Sichuan Agricultural University, Chengdu, China

[3] School of Architecture and Urban and Rural Planning, Sichuan Agricultural University, Chengdu, China

[4] School of Mathematics and Computer Application Technology, Jining University, Jining, China

* Corresponding Author Email: wenxin_susie@163.com

Abstract. This study uses precipitation, soil moisture and evapotranspiration data as independent variables to predict future soil moisture in the Xilingol grassland of Inner Mongolia, China, while keeping the grazing strategy unchanged. Firstly, the data with anomalous values in the obtained dataset were differenced, and the soil moisture data for different depths were split to build a multi-group ARIMA prediction model. The reliability of the model was determined to obtain the soil moisture in grassland at different depths from 2022 to 2023. Through this study, we have added VAR to the ARIMA model to improve the limitations of predicting future soil moisture based on a single variable and to improve the accuracy of soil moisture prediction, thus providing some theoretical support for the ecological restoration and sustainable development of grasslands.

Keywords: Soil moisture, ARIMA time series, VAR model.

1. Introduction

China is a large grassland country, and in terms of all types of land resources in China, grassland resources cover the largest area, accounting for 40.9% of the country's land area, which is 2.91 times the area of arable land and 1.89 times the area of forests [1]. Grassland is not only a production base for material products, but also has a supporting function, regulating function and cultural service function. Soil moisture is the amount of water contained in the soil and is divided into mass and volume water content. The dry and wet condition of the soil is closely related to the physical and chemical properties of the soil. Monitoring of ecological changes, environmental restoration, water resource dynamics and agricultural production are all closely related to soil moisture content, so monitoring of soil moisture content is of great importance for environmental, agricultural and water resource management [2].

Grazing is the main form of human activity in grassland ecosystems [3]. In this study, soil evaporation data, precipitation data and historical soil moisture data were obtained for the Xilinguole grassland in Inner Mongolia, China, to predict soil moisture at 10cm, 40cm, 100cm and 200cm depths from 2022 to 2023, while keeping the current grazing practices and grazing status unchanged. The data were obtained for all months from January 2012 to March 2022, and a VAR-ARIMA [4] model was developed to predict soil moisture from 2022 to 2023 based on historical soil moisture data. The ARIMA (autoregressive moving average) model was first developed for soil moisture at 10cm, 40cm, 100cm and 200cm depths based on the historical soil moisture data, and the ARIMA model at four different depths was used to predict the soil moisture data for 2022-2023 after the model was tested and analysed to determine a good fit. In addition, the ARIMA model was improved by introducing a VAR (vector autoregressive regression) model, as the ARIMA model only takes into account the influence of single variables, but soil moisture, soil evapotranspiration and precipitation have an interaction [5]. A VAR model was developed for soil moisture, soil evapotranspiration and precipitation at different soil depths, and after testing the model fit, the soil moisture data for 2022-

2023 was predicted. The final results were obtained by combining the ARIMA model and the VAR model.

2. Materials and methods

2.1. Data acquisition and pre-processing

The data used in this study were obtained from the 19th China Postgraduate Mathematical Modelling Competition 2022 in Guangguo, China. The basic data on the soil and climate of the Xilinguole grassland were monitored and provided by specialized institutions, including data on soil moisture, soil evaporation, and the vegetation index NDVI under maintaining the current grazing strategy. (https:/cpipc.acge.org.cn/cw/hp/4)

For the Xilinguole climate 2012-2022 data, after wavelet monitoring, it was found that there were anomalous values in the December 2018 data, which were combined with the analysis results to fill in the linear differences for the December 2012-2022 soil moisture data to ensure the accuracy of the subsequent analysis. And generate soil moisture observation maps for each depth from 2012-2022 (as in Figure 1a-d),10 and 40cm soil moisture have obvious temporal seasonality.

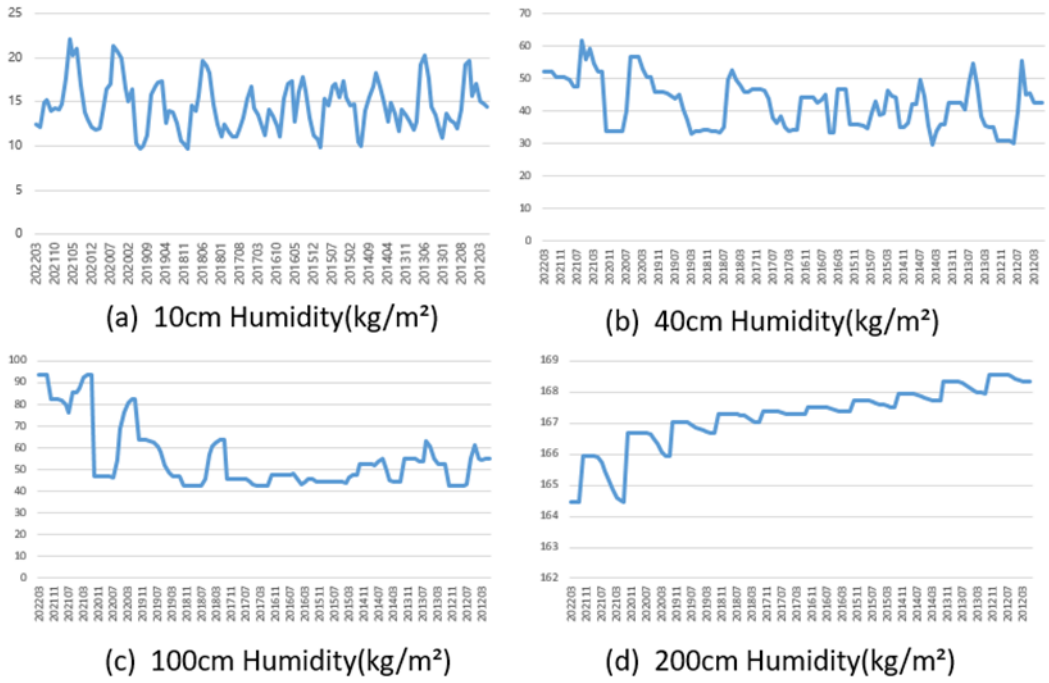


Figure 1. Soil moisture observation maps at different depths from 2012-2022

2.2. Introduction to the method

ARIMA only makes predictions for a single variable of moisture and does not take into account the interaction of soil evaporation and precipitation with soil moisture. For this reason, this paper is improved by using vector autoregression (VAR), a vector autoregressive model proposed by Sims in 1980, which allows three variables, soil moisture, soil evaporation and precipitation, to be put together and predicted as a system in order to make the predictions self-consistent with each other, called a multivariate time series. By plotting the trend of soil moisture over time, it can be used as a reference for the accuracy of the subsequent prediction results for different depths of soil, as shown in Figure 2a-d. It can be seen that soil moisture in the grasslands of Xilinguole League in Inner Mongolia shows an increasing trend at the depths of 10cm, 40cm and 100cm, but a decreasing trend at the depth of 200cm, which indicates that the climate of the Xilinguole League grasslands may be deteriorating and groundwater resources are being lost.

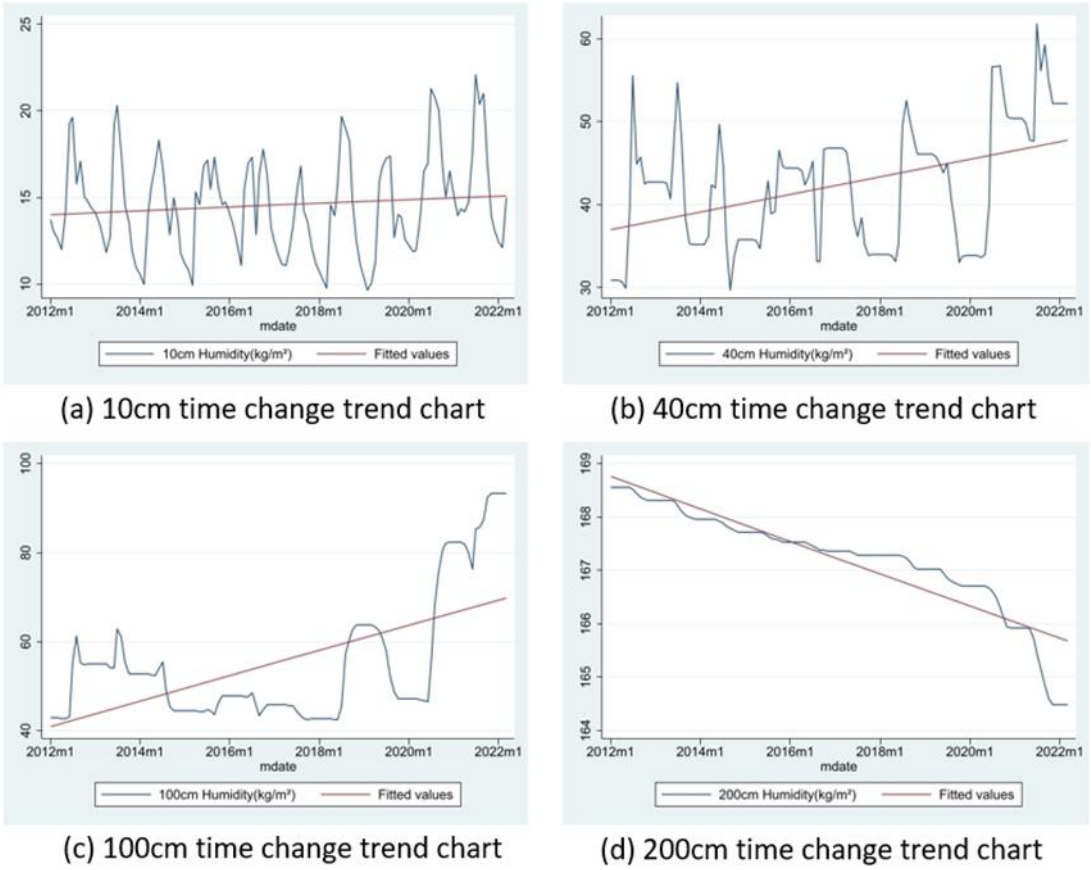


Figure 2. Trend of soil moisture over time at different depths

2.3. Model evaluation indicators

In this paper, Granger causality test, smoothness test and residual test are chosen as indicators for the evaluation of the model.

2.3.1 Granger causality test

The Granger causality test statistically tests the causality of variables over a time series, providing evidence for the analysis of potential causal relationships between variables[6], and it is a prerequisite for whether an impulse function can be established.

In a VAR model, the Granger test of causality is not what is commonly referred to as causality, but rather whether what happens first has some effect on what happens later, or whether a variable can be used to improve the predictive power of other variables of interest. Therefore, "Granger causality is essentially a "predictive" relationship. It is essentially a consideration of whether the lag of one variable can be added to the equation of other variables. When one variable is indeed influenced by the lags of the other variable, the two variables are said to have Granger causality.

2.3.2 Stability test

Before a VAR model can be fitted, the variables need to be tested for stationarity. A VAR model can only be fitted if the fitted endogenous variables are all stationary or single integer of the same order. However, the two variables need to be logarithmic prior to the smoothness test to eliminate the effect of heteroskedasticity in the time series.

2.3.3 Residual test

The established models are tested by residual tests to determine whether the established VAR models are better, to check whether the perturbation terms of the models obey a normal distribution and whether there is serial autocorrelation.

3. Model building and solving

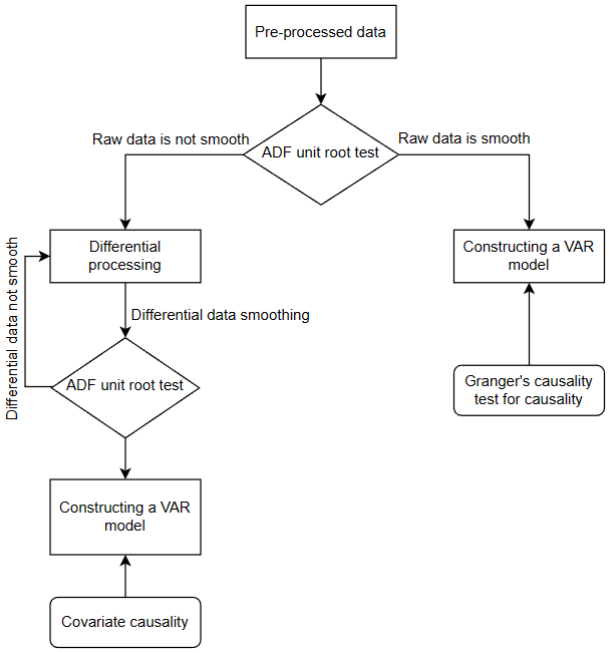


Figure 3. VAR model flow chart

The data were first pre-processed, followed by an ADF unit root test. If the original data are smooth, a VAR model is constructed. If the original data is not smooth, the differencing process is performed until the differenced data is smooth, and then the VAR model is constructed [7]. Finally, the model was subjected to Granger causality test for causality [8]. the flow chart of the VAR model is shown in Figure 3.

3.1. Model building

Assuming that the time series variables of soil moisture, soil evapotranspiration and precipitation at 10 cm are y_{1t} 、 y_{2t} and y_{3t} , respectively as the explanatory variables of the three regression equations, and the explanatory variables are the p-order lagged values of these three variables, a ternary VAR(p) system can be established as follows.

$$\begin{cases} y_{1t} = \beta_{10} + \beta_{11} y_{1,t-1} + \ldots + \beta_{1p} y_{1,t-p} + \ldots + \gamma_{1p} y_{2,t-p} + \varepsilon_{1t} \\ y_{2t} = \beta_{20} + \beta_{21} y_{2,t-1} + \ldots + \beta_{2p} y_{2,t-p} + \ldots + \gamma_{2p} y_{3,t-p} + \varepsilon_{2t} \\ y_{3t} = \beta_{30} + \beta_{31} y_{3,t-1} + \ldots + \beta_{3p} y_{3,t-p} + \ldots + \gamma_{3p} y_{1,t-p} + \varepsilon_{3t} \end{cases} \quad (1)$$

The VAR equation sets for soil moisture, soil evaporation and precipitation for 40cm, 100cm and 200cm were obtained in the same way. For ease of analysis, the VAR equations constructed for soil moisture at 10cm, 40cm, 100cm and 200cm were named equation set 1, equation set 2, equation set 3 and equation set 4 respectively. and analysed using Stata 17.0.

3.1.1 Unit root test

The purpose of the test is to determine if the data is suitable for the VAR and the time series data is smooth before the VAR model can be used. If the time series data is not smooth, it needs to be differenced or logarithmically processed to make the data smooth. The results are shown in Table 1.

Unit root test for equation set 1: a Z value of 0.001 for soil evapotranspiration and 0.000 for precipitation indicates that the series data for both variables is smooth. a Z value of 0.000 for the time series data for soil moisture at 10 cm indicates that the series data for this variable is smooth.

Unit root test for equation set 2: The time series of soil moisture at 40 cm had a Z value of 0.009, indicating that the series of this variable was stable.

Unit root test for equation set 3: The time series of soil moisture at 100 cm has a Z value of 0.973, indicating that the series is not stable. The Z-value of the unit root test for soil moisture at 100cm was 0.000, indicating that the series data was stable and passed the unit root test.

Unit root test for equation group 4: The time series data of soil moisture at 200 cm has a Z value of 1.000, indicating that the series data of this variable is non-stationary. The result of the unit root test after differencing was 0.018, indicating that the series data passed the unit root test and was stable.

Table 1. Table of unit root test results

Soil moisture	Z-value	results	Differential processing results
System of equations 1	0.000	Stable	-
System of equations 1	0.009	Stable	-
System of equations 1	0.973	Unstable	0.000
System of equations 1	1.000	Unstable	0.018

3.1.2 Determination of the lag order

In VAR modelling, it is important to determine the lagged order of the variables, and metrics such as LR, FPE, AIC, HQIC and SBIC are provided in Stata. LR is the likelihood ratio test, which is a likelihood ratio test of the joint significance of the last order coefficients. fpe denotes Akaike's Prediction Error. AIC, HQIC and SBIC are information criterion indicators. The number of lags used in the model was determined by looking at the indicators marked with an asterisk in the table. The best choice for equation set 1 is to choose a lag of order 2. Similarly, equation group 2 was chosen with a 4th order lag, equation group 3 with a 4th order lag and equation group 4 with a 4th order lag. The screenshots of the results are omitted due to space limitations.

3.1.3 Estimating the VAR model

As the VAR model includes many parameters whose meaning is difficult to interpret, the analysis is mainly carried out through the impulse corresponding function. In practice, the order of variables in the OIRF (orthogonalised impulse response function) is mainly determined by crossing the cross-correlogram of two variables and Granger causality tests.

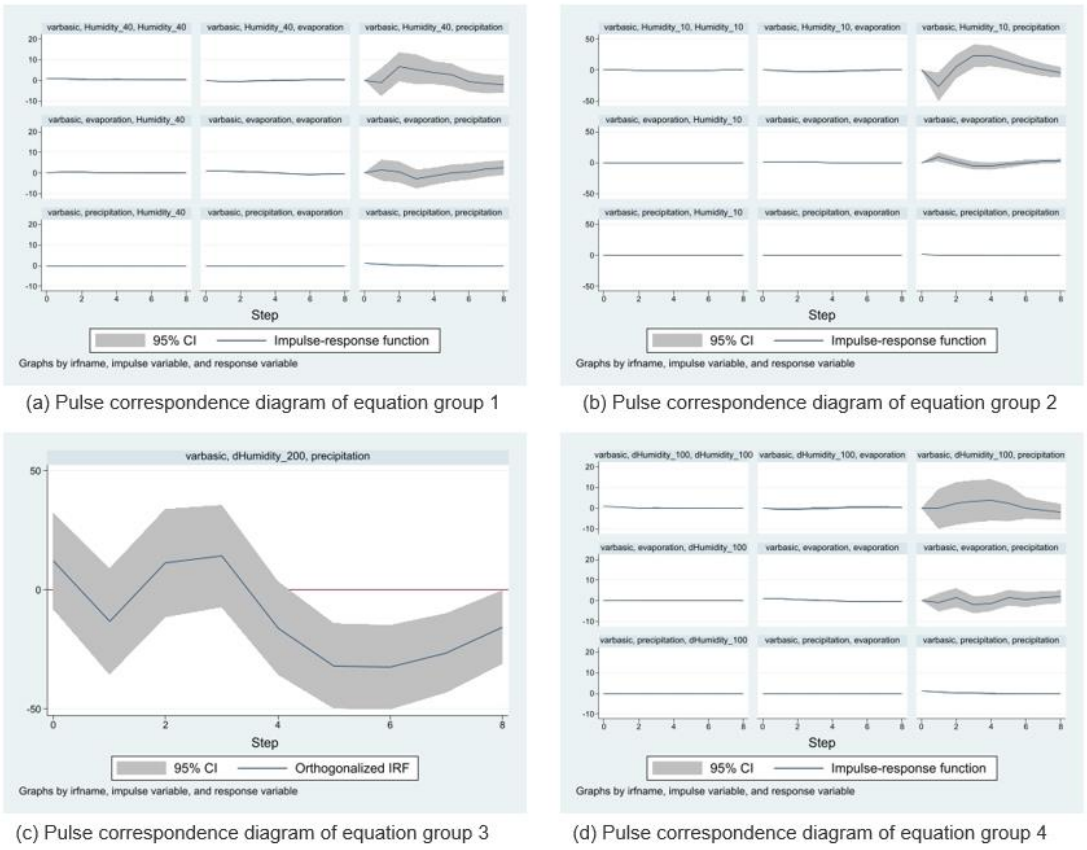
1) Cross-correlation diagram

The cross-correlation function for variables y_t and x_t is defined as follows :

$$\rho_{yx}(k) \equiv corr(y_t, x_{t+k}) \tag{2}$$

That is, the correlation coefficient between y_t and x_{t+k}, $\rho_{yx}(k)$ as a function of k, and draw a graph, you can get the cross-correlation graph, through the cross-correlation graph, you can see that the k value that makes $|\rho_{yx}(k)|$ the largest, that is, k*, if k* > 0, y_t and the future x_{t+k} most relevant, the variables are sorted as y_t , x_t ; conversely, if k* < 0, y_t and the future x_{t+k} most relevant variables are sorted as x_t , y_t .

The orthogonalised pulses for equation set 1 are shown accordingly in Figure 4, with the first row of vignettes depicting the dynamic effects of soil evaporation on soil evapotranspiration, soil moisture at 10cm and precipitation respectively, and the second and third row of vignettes following on from this. It can be seen that soil evaporation has a significant effect on precipitation, soil moisture also has a significant effect on precipitation and the remaining variables have no significant effect. The impulse response plots for equation set 2, equation set 3 and equation set 4 are shown in Figure 4a-d respectively. As can be seen from the plots, the impulse response of soil moisture at 40cm depth and 100cm depth is generally consistent with that at 10cm depth, while at 200cm depth, only moisture has a significant effect on precipitation, with the remaining effects being insignificant.



(a) Pulse correspondence diagram of equation group 1

(b) Pulse correspondence diagram of equation group 2

(c) Pulse correspondence diagram of equation group 3

(d) Pulse correspondence diagram of equation group 4

Figure 4. Corresponding graphs of pulses for different sets of equations

2) Granger causality test

Granger (1969) proposed the Granger causal model [9] based on the idea that past values of x can help predict future values of y if x is a cause of y, but y is not a cause of x.

$$y_t = \gamma + \sum_{m=1}^{p} \alpha_m y_{i,t-m} + \sum_{m=1}^{p} \beta_m x_{i,t-m} + u_i + \varepsilon_{it} \tag{3}$$

The results of the Granger causality test for equation set 1 are shown in Figure 5-a. The upper part of the figure shows that in the equation with soil moisture at 10cm depth as the explanatory variable, if the joint significance of the coefficients for soil evaporation is tested, its chi-square statistic is 19.862 and the corresponding p-value is 0.000, so it can be assumed that soil evaporation is the Granger cause of soil moisture at 10cm depth. Similarly, if the joint significance of the coefficients for precipitation is tested, the chi-square statistic is 0.756 and the corresponding p-value is 0.685, so it can be assumed that precipitation is not the Granger cause of soil moisture at 10 cm depth. However, the overall chi-square statistic is 21.585 and the corresponding p-value is 0.000, so it can be assumed that soil moisture at 10cm depth is influenced by soil evaporation and precipitation. By analogy, the results of the Granger test can be found in the middle and next steps of the graph for soil evaporation and precipitation as the explained variables and the remaining two variables as the explanatory variables, respectively. The test results show that there is an interaction between soil moisture, soil evapotranspiration and precipitation at different depths. The results of the Granger causality tests for equation set 2, equation set 3 and equation set 4 are shown in Figure 5-b to 5-d respectively, and the test results are basically consistent with those of equation set 1.

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
Humidity_10	evaporation	19.862	2	0.000
Humidity_10	precipitation	.75581	2	0.685
Humidity_10	ALL	21.585	4	0.000
evaporation	Humidity_10	23.23	2	0.000
evaporation	precipitation	2.3145	2	0.314
evaporation	ALL	26.06	4	0.000
precipitation	Humidity_10	14.623	2	0.001
precipitation	evaporation	10.728	2	0.005
precipitation	ALL	15.054	4	0.005

(a) Granger causality diagram of equation group 1

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
Humidity_40	evaporation	22.787	4	0.000
Humidity_40	precipitation	1.8984	4	0.754
Humidity_40	ALL	23.471	8	0.003
evaporation	Humidity_40	21.632	4	0.000
evaporation	precipitation	2.6442	4	0.619
evaporation	ALL	26.92	8	0.001
precipitation	Humidity_40	5.7643	4	0.217
precipitation	evaporation	5.9269	4	0.205
precipitation	ALL	9.2864	8	0.319

(b) Granger causality diagram of equation group 2

Granger causality Wald tests

Equation	Excluded	chi2	df	Prob > chi2
dHumidity_100	evaporation	43.725	4	0.000
dHumidity_100	precipitation	1.4949	4	0.828
dHumidity_100	ALL	45.826	8	0.000
evaporation	dHumidity_100	9.7806	4	0.044
evaporation	precipitation	3.7993	4	0.434
evaporation	ALL	14.276	8	0.075
precipitation	dHumidity_100	.65946	4	0.956
precipitation	evaporation	3.1925	4	0.526
precipitation	ALL	4.0084	8	0.856

(c) Granger causality diagram of equation group 3

Granger causality Wald tests

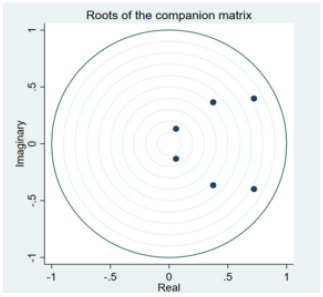
Equation	Excluded	chi2	df	Prob > chi2
dHumidity_200	evaporation	26.426	4	0.000
dHumidity_200	precipitation	.60644	4	0.962
dHumidity_200	ALL	26.989	8	0.001
evaporation	dHumidity_200	4.5042	4	0.342
evaporation	precipitation	2.2655	4	0.687
evaporation	ALL	8.8138	8	0.358
precipitation	dHumidity_200	25.275	4	0.000
precipitation	evaporation	1.0303	4	0.905
precipitation	ALL	29.319	8	0.000

(d) Granger causality diagram of equation group 4

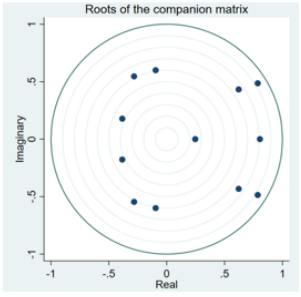
Figure 5. Granger causality plots for different sets of equations

3.1.4 Stability tests

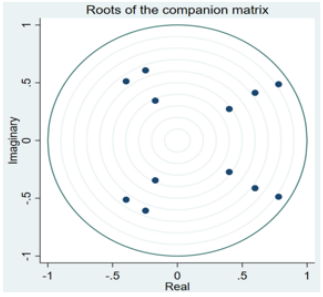
This step is to check whether the VAR system is smooth (as a smooth process). The smoothness test of equation set 1 is shown in Figure 6-a, which shows that all the eigenvalues are within the unit circle, so the constructed VAR system model is smooth. Similarly, the results of the smoothness tests at 40cm, 100cm and 200cm depths can be obtained, as shown in Figure 6-b to 6-d respectively, which show that the constructed VAR models are stable.



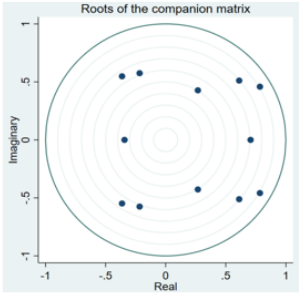
(a) Equation 1 smoothness test



(b) Equation 2 smoothness test



(c) Equation 3 smoothness test



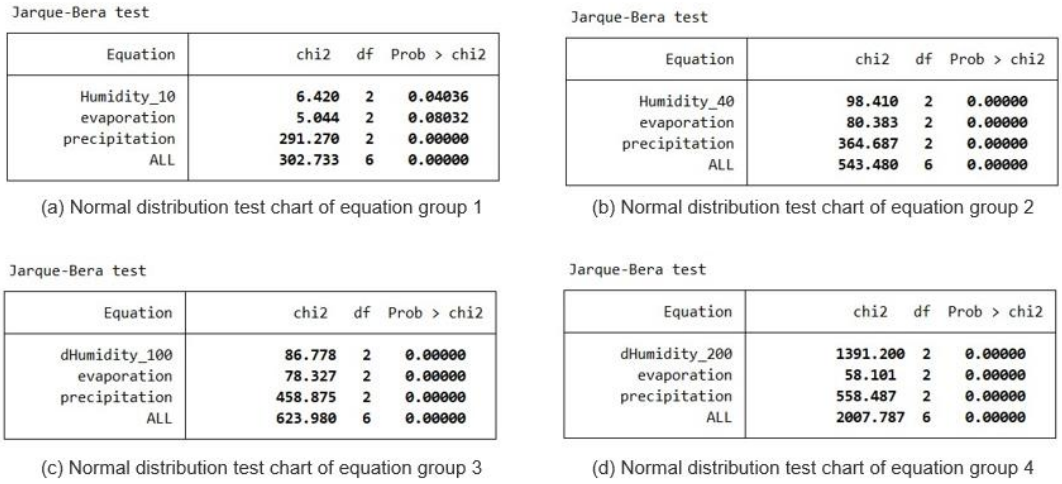
(d) Equation 4 smoothness test

Figure 6. Smoothness tests for different systems of equations

3.1.5 Residual tests

The established models were tested by residual tests to determine whether the established VAR models were better, in order to test whether the perturbation terms of the models obeyed a normal distribution and whether there was serial autocorrelation.

The results of the normal distribution tests for equation set 1, equation set 2, equation set 3 and equation set 4 are shown in Figure 7-a to 7-d. The results show that the model perturbation terms reject the original hypothesis at the 1% significance level and the model perturbation terms are considered not to obey a normal distribution.



Jarque-Bera test

Equation	chi2	df	Prob > chi2
Humidity_10	6.420	2	0.04036
evaporation	5.044	2	0.08032
precipitation	291.270	2	0.00000
ALL	302.733	6	0.00000

(a) Normal distribution test chart of equation group 1

Jarque-Bera test

Equation	chi2	df	Prob > chi2
Humidity_40	98.410	2	0.00000
evaporation	80.383	2	0.00000
precipitation	364.687	2	0.00000
ALL	543.480	6	0.00000

(b) Normal distribution test chart of equation group 2

Jarque-Bera test

Equation	chi2	df	Prob > chi2
dHumidity_100	86.778	2	0.00000
evaporation	78.327	2	0.00000
precipitation	458.875	2	0.00000
ALL	623.980	6	0.00000

(c) Normal distribution test chart of equation group 3

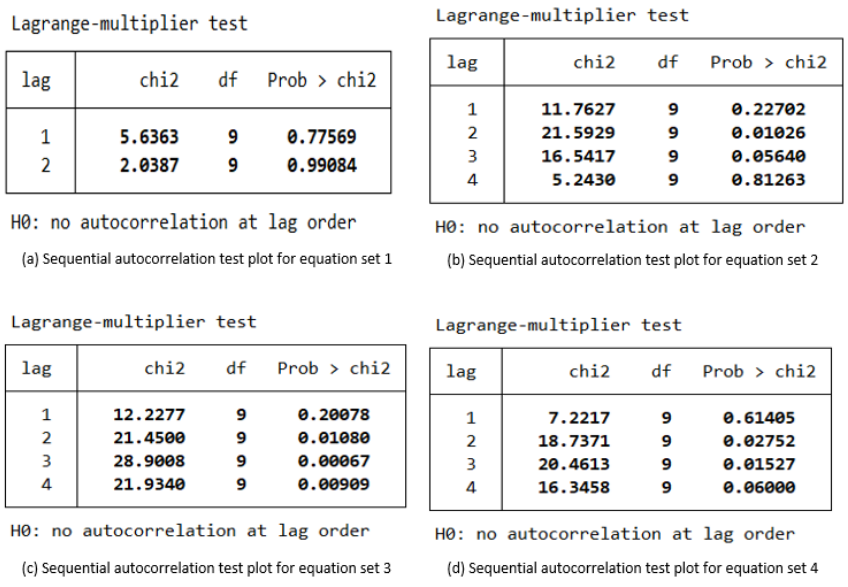
Jarque-Bera test

Equation	chi2	df	Prob > chi2
dHumidity_200	1391.200	2	0.00000
evaporation	58.101	2	0.00000
precipitation	558.487	2	0.00000
ALL	2007.787	6	0.00000

(d) Normal distribution test chart of equation group 4

Figure 7. Test plots of normal distribution for different sets of equations

The results of the serial autocorrelation tests for equation set 1, equation set 2, equation set 3 and equation set 4 are shown in Figures 8-a to 8-d. The results show that the model perturbation terms for equation set 1 and equation set 2 reject the original hypothesis, i.e. there is no serial autocorrelation in the perturbation terms. The model perturbation terms of equation group 3 and equation group 4 have serial autocorrelation.



Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	5.6363	9	0.77569
2	2.0387	9	0.99084

H0: no autocorrelation at lag order

(a) Sequential autocorrelation test plot for equation set 1

Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	11.7627	9	0.22702
2	21.5929	9	0.01026
3	16.5417	9	0.05640
4	5.2430	9	0.81263

H0: no autocorrelation at lag order

(b) Sequential autocorrelation test plot for equation set 2

Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	12.2277	9	0.20078
2	21.4500	9	0.01080
3	28.9008	9	0.00067
4	21.9340	9	0.00909

H0: no autocorrelation at lag order

(c) Sequential autocorrelation test plot for equation set 3

Lagrange-multiplier test

lag	chi2	df	Prob > chi2
1	7.2217	9	0.61405
2	18.7371	9	0.02752
3	20.4613	9	0.01527
4	16.3458	9	0.06000

H0: no autocorrelation at lag order

(d) Sequential autocorrelation test plot for equation set 4

Figure 8. Plot of serial autocorrelation tests for different sets of equations

The above tests concluded that the four VAR models constructed fitted well and were suitable for the next step of out-of-sample multi-step forecasting.

3.2. Solving the model

Through the above tests, a VAR model with a good fit was constructed to predict soil moisture at different depths from April 2022 to December 2023 [10]. For the VAR(p) model

$y_t = \tau_0 + \tau_1 y_{t-1} + ... + \tau_p y_{t-p} + \varepsilon_t$, after obtaining parameter estimates, the prediction model for the forward period is

$$y_t = \tau_0 + \tau_1 y_t + ... + \tau_p y_{t-p} + \tau_p y_{t-p+1} \tag{4}$$

where it is assumed that y_t is known, i.e. the current time is period t. In order to predict y_{t+2} , then y_{t+1} can be considered as y_{t+1} .

$$y_{t+2} = \tau_0 + \tau_1 y_{t+1} + \tau_2 y_t + ... + \tau_p y_{t-p+2} \tag{5}$$

Similarly, the prediction y_{t+h} for the forward h period can be calculated . The results of the predictions are shown in Table 2.

Table 2. Var model prediction results table

Year	Month	10cm humidity	40cm humidity	100cm humidity	200cm humidity
2022	4	13.954	37.96	50.128	168.512
	5	12.841	39.892	53.664	168.144
	6	13.242	30.292	46.834	167.841
	7	15.885	39.656	44.462	167.664
	8	15.654	40.836	44.958	167.579
	9	12.297	46.759	45.86	167.371
	10	11.722	34.461	42.648	167.285
	11	13.717	45.993	54.415	167.071
	12	12.649	42.322	54.656	166.84
2023	1	18.15	45.444	48.879	166.653
	2	14.943	51.25	80.494	165.969
	3	13.715	52.161	93.437	164.487
	4	15.885	39.656	44.462	167.664
	5	15.654	40.836	44.958	167.579
	6	12.297	46.759	45.86	167.371
	7	11.722	34.461	42.648	167.285
	8	13.717	45.993	54.415	167.071
	9	12.649	42.322	54.656	166.84
	10	18.15	45.444	48.879	166.653
	11	14.943	51.25	80.494	165.969
	12	13.715	52.161	93.437	164.487

4. Conclusions

The model in this paper can be well applied to known areas of grazing intensity and can predict local soil moisture at different depths. The model can be used to find the economically optimal amount of grazing for the local area while preserving ecological and sustainable development, and serves the function of preventing the emergence of land desertification and soil slumping in a timely manner.

References

[1] Wang Y, Zhang Y, Yu X, et al. Grassland soil moisture fluctuation and its relationship with evapotranspiration[J]. Ecological Indicators, 2021, 131: 108196.

[2] Dao.Guo Study on Distribution Characteristics of Soil Water Content in Xilingol Grassland[D]. Chinese Master's Theses Full-text Database,2018.

[3] DING C, YANG X, Dong Q. Effects of Grazing Patterns on Vegetation, Soil and Microbial Community in Alpine Grassland of Qinghai-Tibetan Plateau[J]. Acta Agrestia Sinica, 2020, 28(1): 159.

[4] Fang-Mei Tseng, Hsiao-Cheng Yu, Gwo-Hsiung Tzeng,Combining neural network model with seasonal time series ARIMA model,Technological Forecasting and Social Change,Volume 69, Issue 1,2002,Pages 71-87,ISSN 0040-1625.

[5] Lütkepohl H. Vector autoregressive models[M]//Handbook of research methods and applications in empirical macroeconomics. Edward Elgar Publishing, 2013: 139-164.

[6] Diks C, Panchenko V. A new statistic and practical guidelines for nonparametric Granger causality testing[J]. Journal of Economic Dynamics and Control, 2006, 30(9-10): 1647-1669.

[7] F. Bashir and H. -L. Wei, "Handling missing data in multivariate time series using a vector autoregressive model based imputation (VAR-IM) algorithm: Part I: VAR-IM algorithm versus traditional methods," 2016 24th Mediterranean Conference on Control and Automation (MED), Athens, Greece, 2016, pp. 611-616, doi: 10.1109/MED.2016.7535976.

[8] Lopez L, Weber S. Testing for Granger causality in panel data[J]. The Stata Journal, 2017, 17(4): 972-984.

[9] Wang H, Li Z, Cao L, et al. Response of NDVI of natural vegetation to climate changes and drought in China[J]. Land, 2021, 10(9): 966.

[10] HE Jun-jie1, WANG Ying-shun1, LI Yun-peng2, WU Ri-na2 Soil Moisture Monitoring with EOS/MODIS VSWI Product in Xilingol[J]. Chinese Journal of Agrometeorology,2013,34(02):243-248.