Statistical analysis and forecasting of cotton yield dynamics in Kashkadarya region of Republic of Uzbekistan

Kudrat Ruzmetov[1*], *Akhtamjon* Faiziev[1], *Salakhiddin* Murodov[1] and *Odina* Kurbonbekova[1]

[1]Tashkent State Agrarian University, University str., 2, 100140, Tashkent province, Uzbekistan

Abstract. There are phenomena that are significant to research because of how they grow and change through time in practically every discipline. One could attempt to direct a process, forecast the future using knowledge of the past, or characterize the distinctive aspects of a series using a finite quantity of information. The techniques used to handle time series are heavily influenced by the techniques created by mathematical statistics for distribution series. The most basic to the most complicated time series analysis techniques exist in statistics today. The article discusses the statistical analysis of a time series, specifically the average yield of cotton in the Kashkadarya region, Uzbekistan, and the Republics, using data from the Central Statistical Office of Uzbekistan from 2001 to 2020. The study involved constructing point and interval estimates for the average cotton yield with a 95% guarantee, identifying different types of trends, and predicting future yields for the region. Through the use of the Durbin-Watson statistical criteria, it was discovered that there is an autocorrelation dependence in the average cotton yield, indicating that the yield for the current year is dependent on yields from past years. The methods used in this study can be applied to further research conducted by students and scientists.

1 Introduction

In almost every field there are phenomena that are important to study in their development and change over time. One can, for example, strive to predict the future on the basis of knowledge of the past, to control a process, to describe the characteristic features of a series on the basis of a limited amount of information. When processing time series, the methods are largely based on the methods developed by mathematical statistics for distribution series. To date, statistics has a variety of methods for analyzing time series from the most elementary to the most complex [1-5].

The statistical analysis and forecasting of cotton yield dynamics involves studying the pattern of change in the yield of cotton over time using statistical methods. This type of analysis is commonly performed using time series analysis, which involves analyzing data collected over a period of time [7-11]. Time series analysis can help identify trends, seasonal patterns, and other patterns in the data, which can be used to make predictions

[*] Corresponding author: Ruzmetovqudrat1967@gmail.com

about future yield. In the context of cotton yield dynamics, statistical analysis may involve collecting data on the average yield of cotton in a particular region or country over a period of time [8-15]. This data can then be analyzed using various statistical methods, including regression analysis, ARIMA modeling, and exponential smoothing. These methods can help identify trends and patterns in the data and make predictions about future yield [7-10, 14-17].

In addition to predicting future yield, statistical analysis of cotton yield dynamics can also be used to identify factors that affect yield. For example, researchers may investigate the impact of weather patterns, irrigation methods, or other agricultural practices on cotton yield. By identifying factors that affect yield, researchers can develop strategies to improve yield and reduce losses [11-14]. Overall, the statistical analysis and forecasting of cotton yield dynamics is an important tool for agricultural researchers and policymakers. By identifying trends, making predictions, and identifying factors that affect yield, this type of analysis can help improve cotton production and ensure food security for communities that depend on cotton as a source of income and sustenance [17-19].

There are three main tasks in the study of time series. The first of them consists in describing the change in the corresponding indicator over time and identifying certain properties of the series under study [7-11]. To do this, they resort to a variety of methods: the calculation of a general indicator of changes in levels over time and the average growth rate; the use of various smoothing filters that reduce fluctuations in levels over time and allow you to more clearly present development trends; selection of curves characterizing this trend; identification of seasonal and other periodic and random fluctuations; measuring the dependence between the members of the series (autocorrelation) [11-15]. The second task of the analysis is to explain the mechanism for changing the levels of the series; to solve it, regression analysis is usually used. In the third, the description of the change in the time series and the explanation of the mechanism for the formation of the series are often used for statistical forecasting, which in most cases comes down to extrapolation of the detected development trends [16-19].

2 Materials and methods

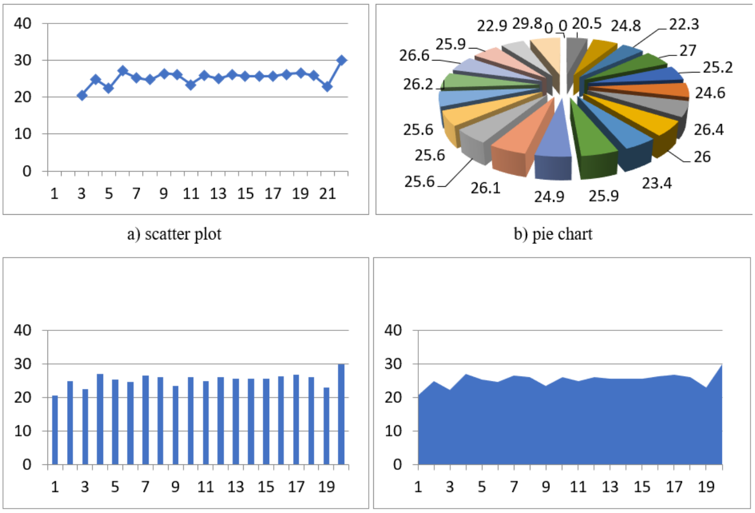
The above mentioned tasks were solved using various methods: 1) The study of the yield of agricultural processes, as a discrete dynamic series and forecasting their yield based on experimental data, play an important role in determining the economic efficiency of farming, dekhkan farms; 2) In this work, the processing and analysis of cotton yield for the observation period 1991-2020 in the Kashkadarya region of the Republic of Uzbekistan was carried out as a discrete time series; 3) Using the methods of statistical analysis of time series, point and interval estimates for the average yield of cotton were constructed , explicit types of trends were determined and the yield was predicted for subsequent years, and various statistical hypotheses were tested [1-7].

In general, the time series $\{y_t,\ t \in T\}$consists of four components: trend; fluctuations relative to the trend; seasonality effect; random component.

3 Results and discussion

The text is describing a research study that focuses on the cotton yields in the Kashkadarya region for the period between 1991 and 2020. The data collected from this study is presented in a table, and the researchers have used this data to analyze $\overline{y_t}$ −the average yield of cotton in the region. The researchers have created four different graphical representations of this data: a scatter plot, a pie chart, a histogram, and a chart with areas.

Each of these graphs displays the data in a different way, which allows for a more comprehensive analysis of the cotton yield dynamics in the region over the given time period (Fig. 1).



a) scatter plot

b) pie chart

c) histogram

d) area chart

Fig. 1. The dynamics of average cotton yield in Kashkadarya region.

The geometric image of the observed data, the coordinate system give grounds, in the first approximation, to assume the hypothesis that the trend part of the process (the general direction of the development of the process) has a linear dependence where the unknown parameters are determined by the least squares method: based on experimental data, solving the following system of normal equations (Table 1):

$$\begin{cases} a_0 T + a_1 \sum t = \sum y_t \\ a_0 \sum t + a_1 \sum t^2 = \sum y_t t \end{cases} \quad (1)$$

Table 1. Calculation of data to determine the trend of the time series.

1	2	3	4	5	6	7
#	Years of observations	y_{t-} q/ha	t	t^2	$y_t\, t$	$y_t\, t^2$
1	2001	20.5	-9	81	-184.5	1660.5
2	2002	24.8	-8	64	-198.4	1587.2
3	2003	22.3	-7	49	-156.1	1092.7
4	2004	27	-6	36	-162	972
5	2005	25.2	-5	25	-126	630
6	2006	24.6	-4	16	-98.4	393.6
7	2007	26.4	-3	9	-79.2	237.6
8	2008	26	-2	4	-52	104

9	2009	23.4	-1	1	-23.4	23.4
10	2010	25.9	0	0	0	0
11	2011	24.9	1	1	24.9	24.9
12	2012	26.1	2	4	52.2	104.4
13	2013	25.6	3	9	76.8	230.4
14	2014	25.6	4	16	102.4	409.6
15	2015	25.6	5	25	128	640
16	2016	26.2	6	36	157.2	943.2
17	2017	26.6	7	49	186.2	1303.4
18	2018	25.9	8	64	207.2	1657.6
19	2019	22.9	9	81	206.1	1854.9
20	2020	29.8	10	100	298	2980
Sum		505.3	10	670	359	16849.4

Solving the system equation (1) and using the calculations in Table 1, we have:

$$\sum y_t = 505{,}3\ q/ha, \qquad a_0 = \frac{1}{\mathrm{T}}\sum y_t = \frac{505{,}3}{20} = 25{,}27\ q/ha,$$
$$a_1 = \frac{1}{\sum t^2}\sum y_t t = \frac{359}{670} = 0{,}54\ q/ha.$$

From here, the equation of the linear trend (trend) of cotton yield in Kashkadarya region found in equations from 1-5.

$$y(t) = 0{,}54t + 25{,}27 \qquad (2)$$

In particular, substituting the value t = 1,2,3 into equation (2) , we find the expected yields of cotton in the Kashkadarya region in 2021 y , will be on average

25,81q/ha, in 2022 - $26{,}35\ q$/ha, and in 2023 - 26.89 q/ha.

With the help of statistical criteria $([1]-[5])$, it was established that in equation (2) $y(t) = a_1 t + a_0$the main hypothesis $H_0 :\ a_1 = 0$ was rejected and an alternative hypothesis $H_1 :\ a_1 \neq 0$ with a significance level was accepted $\alpha = 0{,}05$.

In many observational problems, the observation sample is statistically independent, but in time series they are usually dependent, and the nature of this dependence can be determined by the position of the observations in the sequence. Autocorrelation is a correlation between successive and preceding members of a time series. It was found that the presence of autocorrelation in the average cotton yields in the region was checked and the following one was obtained:

Y $_t$ \u003d ρY $_{t-1}$ + ε$_t$,
g de ρ = Cov (Y $_t$,Y $_{t+1}$) = M $[(Y_t - \overline{y_t})(\ Y_{t+1} - \overline{y_t})]$.

For further research, it was necessary to calculate the following finite differences on the observed data (Table 2).

$$\Delta Y_t = Y_{t+1} - Y_t, \qquad \Delta^2 Y_t = \Delta Y_{t+1} - \Delta Y_t, \qquad \Delta^3 Y_t = \Delta^2 Y_{t+1} - \Delta^2 Y_t$$

The results of the difference in the observed data is given in Table 2.

Table 2. Difference in the given observed data

1	2	3	4	5	6	7	8	9
Years of observations	Y(t) q/ha	Y_t^2	ΔY_t	ΔY_t^2	$\Delta^2 Y_t$	$\Delta^2 Y_t^2$	$\Delta^3 Y_t$	$\Delta^3 Y_t^2$
2001	20.5	420.25	-	-	-	-	-	-
2002	24.8	615.04	4.3	18.49	-	-	-	-
2003	22.3	497.29	-2.5	6.25	-6.8	46.24	-	-
2004	27	729	4.7	22.09	7.2	51.84	14	196
2005	25.2	635.04	-1.8	3.24	-6.5	42.25	-13.7	187.69
2006	24.6	605.16	-0.6	0.36	1.2	1.44	7.7	59.29
2007	26.4	696.96	1.8	3.24	2.4	5.76	1.2	1.44
2008	26	676	-0.4	0.16	-2.2	4.84	-4.6	21.16
2009	23.4	547.56	-2.6	6.76	-2.2	4.84	15	5E-29
2010	25.9	670.81	2.5	6.25	5.1	26.01	7.3	53.29
2011	24.9	620.01	-1	1	-3.5	12.25	-8.6	73.96
2012	26.1	681.21	1.2	1.44	2.2	4.84	5.7	32.49
2013	25.6	655.36	-0.5	0.25	-1.7	2.89	-3.9	15.21
2014	25.6	655.36	0	0	0.5	0.25	2.2	4.84
2015	25.6	655.36	0	0	0	0	-0.5	0.25
2016	26.2	686.44	0.6	0.36	0.6	0.36	0.6	0.36
2017	26.6	707.56	0.4	0.16	-0.2	0.04	-0.8	0.64
2018	25.9	670.81	-0.7	0.49	-1.1	1.21	-0.9	0.81
2019	22.9	524.41	-3	9	-2.3	5.29	-1.2	1.44
2020	29.8	888.04	6.9	47.61	9.9	98.01	12.2	148.84
Sum	505.3	12837.67	9.3	127.15	2.6	308.36	16.7	797.71

Afterwards, there was:

$$V_k = \frac{\sum_{t=k}^{T} \left(\Delta^k Y_t\right)^2}{(T-k) C_{2k}^{k}} \qquad (3)$$

It was reported that the coefficients of variation of differences and found that $V_1 \approx V_2 \approx V_3$. Therefore, first-order finite differences eliminate the linear trend (Table 3).

Table 3. Calculation of data to determine indicators of autocorrection.

1	2	3	4	5	6	7
T	Y_t	$Y_t \cdot Y_{t+1}$	$Y_t \cdot Y_{t+2}$	$Y_t \cdot Y_{t+3}$	$Y_t \cdot Y_{t+4}$	$Y_t \cdot Y_{t+5}$

2001	20.5	-	-	-	-	-
2002	24.8	508.4	-	-	-	-
2003	22.3	553.04	457.15	-	-	-
2004	27	602.1	669.6	553.5	-	-
2005	25.2	680.4	561.96	624.96	516.6	-
2006	24.6	619.92	664.2	548.58	610.08	504.3
2007	26.4	649.44	665.28	712.8	588.72	654.72
2008	26	686.4	639.6	655.2	702	579.8
2009	23.4	608.4	617.76	575.64	589.68	631.8
2010	25.9	606.06	673.4	683.76	637.14	652.68
2011	24.9	644.91	582.66	647.4	657.36	612.54
2012	26.1	649.89	675.99	610.74	678.6	689.04
2013	25.6	668.16	637.44	663.04	599.04	665.6
2014	25.6	655.36	668.16	637.44	663.04	599.04
2015	25.6	655.36	655.36	668.16	637.44	663.04
2016	26.2	670.72	670.72	670.72	683.82	652.38
2017	26.6	696.92	680.96	680.96	680.96	694.26
2018	25.9	688.94	678.58	663.04	663.04	663.04
2019	22.9	593.11	609.14	599.98	586.24	586.24
2020	29.8	682.42	771.82	792.68	780.76	762.88
Sum	505.3	12119.95	11579.78	10988.6	10274.52	9611.36

Using Table 3, the formulas from the literature [1-5] determine the values of the autocorrelation coefficients R_L $out\ of$ $L = 1,2,3,4,5$ (where: L_a, time shift, i.e. the time interval of one phenomenon lagging behind the other associated with it):

$$R_l = \frac{\sum_{t=1}^{N-L} Y_t Y_{t+L} - \frac{\sum_{t=1}^{N-L} Y_t \sum_{t=L+1}^{N} Y_t}{N-L}}{\sqrt{\left[\sum_{t=1}^{N-L} Y_t{}^2 - \frac{\left(\sum_{t=1}^{N-L} Y_t\right)^2}{N-L}\right]}\left[\sum_{t=L+1}^{N} Y_t{}^2 - \frac{\left(\sum_{t=L+1}^{N} Y_t\right)^2}{N-L}\right]} \qquad (4)$$

The difference of the value $R_L$from zero gives reason to believe that there is a significant autocorrelation between the yield of cotton. It was reported that it was indented to check the hypothesis of the existence of an autocorrelation dependence between the yield of cotton using the Durbin -Watson criterion:

$$d = \frac{\sum_{t=1}^{T-1}(Y_{t+1} - Y_t)^2}{\sum_{t=1}^{T} Y_t{}^2} \qquad (5)$$

Therefore, the Durbin -Watson criterion of 95% also proves with a guarantee that the average cotton yield in the region has an autocorrelation dependence $Y_t = \rho Y_{t-1} + \varepsilon_t$. Consequently, the yield of cotton in the region this year depends on the yield of past years. Furthermore, in this research, Testing the statistical hypothesis of normality, $\bar{y}_t$ −the average cotton yield in Kashkadarya region $([1] - [5])$:

$$\mathrm{H}_0 : P(\bar{y}_t < x) = F_{\mathrm{a},\sigma}(\mathrm{x}), \qquad \mathrm{H}_1\text{: } P(\bar{y}_t < x) \neq F_{\mathrm{a},\sigma}(\mathrm{x})$$

According to the results of the calculation, significance level was accepted$\alpha = 0{,}05\ \text{cm}\ (\text{Table 4})$.

Then, using the following formulas, we construct interval estimates for the average cotton yield:

$$\bar{Y}_{T+i} - t(T-2;\alpha)\bar{\sigma}_y \leq a_0 + a_1(T+i) \leq \bar{Y}_{T+i} + t(T-2;\alpha)\bar{\sigma}_y (6)$$

$$\text{Where } \bar{\sigma}_y = \bar{\sigma}\left[\frac{\frac{1}{T}+\left(\frac{T-1}{2}+i\right)^2}{\sum_{i=1}^{T}(t-\bar{t})^2}\right]^{0.5}$$

The results of the statical analysis depicted that average cotton yield was 25.27, followed by dispersion (3.75), standard deviation (1.94), coefficient of variation (7.68 %), asymmetry (-0.40). Noteworthy, statistical hypothesis testing was calculated, and accordingly, it was 95% guarantee of the hypothesis, which means that $H_0$was accepted (Table 4).

Table 4. Estimation of the main parameters of the dynamic series.

Selected characteristics	**Estimates of sample characteristics**
Average cotton yield $\bar{y}_T$q/ha	25.27
Dispersion	3.75
Standard deviation σ_T	1.94
Coefficient of variation v(%)	7.68 %
Asymmetry A_{ς}	-0.40
excess $E_{K\ \varsigma}$	1.98
Error of the mean $\bar{y}_T$,m_y	$m_y = \frac{\sigma_y}{\sqrt{n}} = 0{,}43$
marginal error m'_y	M' y \u003d tm y \ u003d 2.09 · 0.43 \u003d 0.9 0
Standard deviation error σ_T	$m_{\sigma} = \frac{\sigma}{\sqrt{2n}} = \frac{1{,}94}{6{,}32} = 0{,}31$
Interval evaluation (95%)$\bar{y}_T \pm tm_y$ for cotton yield	$\bar{y}_T \pm tm_y = 25.27 \pm 0.90$ (2 4.37 ; 26 . 17) q / ha
Statistical hypothesis testing $H_0 : P(\bar{y}_t < x) = \Phi_{a,\sigma}(x)$	95% guarantee of the hypothesis $H_0$is accepted

4 Conclusions

Based on the above statistical analyzes, the dynamics $\bar{y}_t$ —of cotton yields in the Kashkadarya region in the Republic of Uzbekistan as a discrete time series with reliability $\gamma = 0{,}95$(Table-4), the following conclusions can be drawn:

1. point and interval statistical estimates for $\bar{y}_t$ − average cotton yield in Kashkadarya region were constructed. In particular, the average $\bar{y}_t$ −cotton yield in Kashkadarya region will be with a 95% guarantee, accounted for 24.37 and 26.17 q/ha;

2. Explicit types of the trend were determined and its linearity was established $y(t) = 0{,}54t + 25{,}27$;

3. Using the Durbin -Watson criterion, it was found that the average cotton yield in the region has an autocorrelation dependence $Y_t = \rho Y_{t-1} + \varepsilon_t$, where $\rho = \text{Cov}(Y_t, Y_{t+1}) = M[(Y_t - \bar{y}_t)(Y_{t+1} - \bar{y}_t)]$. In general, it has been proven that the dynamics of the average cotton yield in the Kashkadarya region forms a non-stationary time series.

References

1. Buriev Kh.Ch., Fayziev A.A., Nishanova A. (2021). Statistical analysis and forecasting Dynamics of harvest quality of gourds *Journal, Bulletin of Agrarian Science of Uzbekistan* **1** (85), pp. 47-52.
2. Tukhtaboev, A. A., Turaev, F., Khudayarov, B. A., Esanov, E., & Ruzmetov, K. (2020, December). Vibrations of a viscoelastic dam–plate of a hydro-technical structure under seismic load. I*n IOP Conference Series: Earth and Environmental Science* **614** (1) pp. 012051. Doi: 10.1088/1755-1315/614/1/012051
3. Pathan, K. A., Dabeer, P. S., & Khan, S. A. (2020). Enlarge duct length optimization for suddenly expanded flows *Advances in Aircraft and Spacecraft Science* **7**(3) pp. 203-214. Doi: 10.12989/aas.2020.7.3.203
4. Khudayarov, B., Ruzmetov, K., Turaev, F., Vaxobov, V., Hidoyatova, M., Mirzaev, S., & Abdikarimov, R. (2020). Numerical modeling of nonlinear vibrations of viscoelastic shallow shells *Engineering Solid Mechanics* **8**(3) pp. 199-204. Doi: 10.5267/j.esm.2020.1.004
5. Kurbonov, N., & Aminov, S. (2020). Computer modeling of filtration processes with piston extrusion *In Journal of Physics: Conference Series* **1441** (1) pp. 012147. Doi: 10.1088/1742-6596/1441/1/012147
6. Gulomov, O. K., Khudayarov, B. A., Turaev, F. Z., & Ruzmetov, K. S. (2021). Quadratic forms related to the voronoi⇔ s domain faces of the second perfect form in seven variables *Dynamics of Continuous, Discrete and Impulsive Systems Series B: Applications and Algorithms* **28**(1) pp. 15-23.
7. Rakhmanov, S., Turgunov, T. T., Kusharov, Z. K., & Mengnorov, A. A. (2020, December). Econometric methods for solving problems of analysis and forecasting dynamics of yield of agricultural crops. In *IOP Conference Series: Earth and Environmental Science* (Vol. 614, No. 1, p. 012165). IOP Publishing.
8. Papageorgiou, E. I., Markinos, A. T., & Gemtos, T. A. (2011). Fuzzy cognitive map based approach for predicting yield in cotton crop production as a basis for decision support system in precision agriculture application. *Applied Soft Computing*, *11*(4), 3643-3657.
9. Zhao, D., Reddy, K. R., Kakani, V. G., Read, J. J., & Koti, S. (2007). Canopy reflectance in cotton for growth assessment and lint yield prediction. *European Journal of Agronomy*, *26*(3), 335-344.
10. Thorp, K. R., Ale, S., Bange, M. P., Barnes, E. M., Hoogenboom, G., Lascano, R. J., ... & White, J. W. (2014). Development and application of process-based simulation models for cotton production: A review of past, present, and future directions. *Journal of Cotton Science*, *18*(1), 10-47.
11. Chen, S., Jiang, T., Ma, H., He, C., Xu, F., Malone, R. W., ... & He, J. (2020). Dynamic within-season irrigation scheduling for maize production in Northwest China: A Method Based on Weather Data Fusion and yield prediction by DSSAT. *Agricultural and Forest Meteorology*, *285*, 107928.
12. Zhu, Y., Goodwin, B. K., & Ghosh, S. K. (2011). Modeling yield risk under technological change: Dynamic yield distributions and the US crop insurance program. *Journal of Agricultural and Resource Economics*, 192-210.
13. Livieris, I. E., Dafnis, S. D., Papadopoulos, G. K., & Kalivas, D. P. (2020). A multiple-input neural network model for predicting cotton production quantity: a case study. *Algorithms*, *13*(11), 273.

14. Aidarova, A., Uskenov, M., Zhakeshova, A., Dosmuratova, E., & Kulanova, D. (2016). The economic analysis and prerequisites for creation of a cotton and textile cluster in the republic of kazakhstan. *Indian Journal of Science and Technology*, *9*(5), 1-5.
15. Darekar, A., & Reddy, A. A. (2017). Cotton price forecasting in major producing states. *Economic Affairs*, *62*(3), 373-378.
16. Xu, W., Chen, P., Zhan, Y., Chen, S., Zhang, L., & Lan, Y. (2021). Cotton yield estimation model based on machine learning using time series UAV remote sensing data. *International Journal of Applied Earth Observation and Geoinformation*, *104*, 102511.
17. Hussein, F., Janat, M., & Yakoub, A. (2011). Simulating cotton yield response to deficit irrigation with the FAO AquaCrop model. *Spanish Journal of Agricultural Research*, *9*(4), 1319-1330.
18. Dhaliwal, J. K., Panday, D., Saha, D., Lee, J., Jagadamma, S., Schaeffer, S., & Mengistu, A. (2022). Predicting and interpreting cotton yield and its determinants under long-term conservation management practices using machine learning. *Computers and Electronics in Agriculture*, *199*, 107107.
19. Aparecido, L. E. D. O., Meneses, K. C. D., Rolim de Souza, G., Carvalho, M. J. N., Pereira, W. B. S., da Silva, P. A., ... & de Moraes, J. R. D. S. C. (2022). Algorithms for forecasting cotton yield based on climatic parameters in Brazil. *Archives of Agronomy and Soil Science*, *68*(7), 984-1001.