AGRI MAGAZINE
(International E-Magazine for Agricultural Articles)
Volume: 03, Issue: 10 (October, 2026)
Available online at http://www.agrimagazine.in
©Agri Magazine, ISSN: 3048-8656

Symbolic Data Analysis for Agricultural Time Series: Methods and Applications

Prabhat Kumar, Santosha Rathod, Prashantkumar. S. Hanjagi, *Sushma. M. Awaji, Suraj Gund, Sudhir Kumar Mishra and R N Singh
ICAR-National Institute of Abiotic Stress Management, Baramati, Maharashtra, India
***Corresponding Author's email: sma1624@gmail.com**

Agriculture is strongly influenced by changing environmental conditions, including temperature, rainfall, soil moisture, and humidity. Understanding how these variables change over time is essential for managing crops, assessing yields, and promoting sustainable farming. Observation stations, remote-sensing systems, agricultural experiments and field sensors produce large quantities of data. The trick is extracting useful information from these observations. Traditional statistical methods often reduce observations to a single value (e.g. the mean). An average can be helpful, but it can mask important differences in variability and distribution. For example, two growing seasons may have the same average temperature but vary in the frequency of extreme heat events. Symbolic Data Analysis (SDA) extends traditional data analysis by describing observations with more complex structures such as intervals, distributions and histograms. These representations allow researchers to study variation within groups or periods, rather than relying only on individual numerical values (Billard & Diday, 2006; Beranger et al., 2023). Two related approaches can be useful: symbolic analysis of time series, which transforms numerical sequences into categorical states to explore temporal patterns, and symbolic regression, which looks for mathematical expressions that describe relationships between variables. These methods can be used together in agricultural research, but they have different objectives and should not be considered interchangeable.

Understanding Symbolic Data Analysis

Interval-valued data

Interval-valued data represent an observation through a lower and an upper bound. For example, if daily temperatures during a particular period range from 18°C to 32°C, the interval can be written as

$$X_i = [a_i, b_i], \quad a_i \le b_i, \tag{1}$$

where a_i and b_i denote the lower and upper bounds of the i-th interval-valued observation, respectively.

The midpoint and range are calculated as

$$m_i = \frac{a_i + b_i}{2}, \tag{2}$$

and

$$r_i = b_i - a_i. \tag{3}$$

For the interval [18,32], the midpoint is 25°C, and the range is 14°C. The midpoint indicates the centre of the interval, whereas the range describes its width.

In agriculture, intervals can represent temperature extremes, rainfall ranges, soil-moisture variation, or variation in crop growth measurements. However, an interval alone does not

indicate how frequently different values occur within it. SDA also includes richer representations, such as histogram-valued data, that retain information about distributions (Billard & Diday, 2006).

Histogram-valued data

A histogram describes how observations are distributed across a set of intervals. A histogram-valued observation can be represented as

$$H_i = \{(I_k, p_{ik})\}_{k=1}^{K}, \quad (4)$$

where I_k is the k-th interval, p_{ik} is the proportion of observations assigned to that interval for the i-th unit, and K is the total number of intervals.

When the p_{ik} values represent proportions, they satisfy

$$p_{ik} \geq 0, \quad \sum_{k=1}^{K} p_{ik} = 1. \quad (5)$$

For example, a histogram could describe the distribution of daily temperatures during a crop's flowering stage. It may reveal whether temperatures were concentrated in a moderate range or whether a substantial proportion of days experienced high temperatures.

This representation is useful when the distribution of environmental conditions matters in addition to their average. The intervals and bin frequencies should be selected consistently so that comparisons between locations or seasons are meaningful.

Comparing distributions

Researchers may need to compare the distributions of rainfall, temperature, or soil moisture across locations or growing seasons. One established measure is the Wasserstein distance, which belongs to the family of distances used in optimal transport theory (Villani, 2009).

For two univariate probability distributions F and G with finite second moments, the squared Wasserstein-2 distance is

$$W_2^2(F, G) = \int_0^1 [Q_F(u) - Q_G(u)]^2 \, du, \quad (6)$$

where $Q_F(u)$ and $Q_G(u)$ are the quantile functions of the two distributions, and $u \in (0,1)$ is the probability level.

In simple terms, this equation compares corresponding quantiles of the distributions. A smaller distance indicates that the distributions are more similar under this measure, while a larger distance indicates greater differences. For agricultural research, it can help compare temperature distributions between seasons or rainfall distributions between regions.

Symbolic Time Series Analysis in Agriculture

A time series consists of observations collected sequentially, such as daily rainfall, monthly temperature, or soil moisture measured by sensors. Symbolic time series analysis transforms numerical observations into a sequence of discrete states, making it possible to examine transitions and recurring patterns (Piccardi, 2004).

3Converting numerical observations into symbols

Consider a daily rainfall record. A researcher might define three states:

- D: dry conditions
- M: moderate rainfall
- H: heavy rainfall

Thresholds are selected to assign each observation to a state. A hypothetical rainfall sequence of

$$(12,\ 0,\ 28,\ 35,\ 4,\ 18) \quad (7)$$

could, under an appropriate classification rule, be converted into

$$(\mathrm{M},\ \mathrm{D},\ \mathrm{H},\ \mathrm{H},\ \mathrm{D},\ \mathrm{M}). \quad (8)$$

The numerical thresholds defining each state must be specified in an actual study. They should reflect the research question, local climate, crop requirements, and observation scale. Different thresholds may produce different symbolic patterns, so their influence should be evaluated.

Transition probabilities and Markov models

After converting observations into symbols, researchers can investigate how often one state follows another. The transition probability from state i to state j is

$$p_{ij} = P(S_{t+1} = j \mid S_t = i), \tag{9}$$

where S_t is the symbolic state at time t.

For observed data, this probability can be estimated using transition counts:

$$\hat{p}_{ij} = \frac{N_{ij}}{\sum_{l=1}^{K} N_{il}}, \tag{10}$$

where N_{ij} is the number of observed transitions from state i to state j, and $\hat{p}_{ij}$ is the estimated transition probability. This estimate is defined when the denominator is greater than zero.

The transition probabilities can be arranged into a $K \times K$ matrix:

$$P = [p_{ij}]_{K\times K}, \quad \sum_{j=1}^{K} p_{ij} = 1, \quad i = 1, \ldots, K. \tag{11}$$

If $\boldsymbol{\pi}_t$ is a row vector describing the probabilities of occupying each state at time t, a first-order Markov model gives

$$\boldsymbol{\pi}_{t+1} = \boldsymbol{\pi}_t P. \tag{12}$$

This equation describes the next-step state probabilities under the assumption that the current state is sufficient to model the next transition. For example, it can help investigate whether dry conditions tend to persist or whether high-temperature states frequently follow moderate-temperature states. The assumption and predictive performance should be tested before the model is used for decision-making.

Entropy and recurring patterns

Shannon entropy measures uncertainty in a probability distribution (Shannon, 1948). For a symbolic sequence with K possible states, it can be expressed as

$$H(S) = -\sum_{i=1}^{K} p_i \log p_i, \tag{13}$$

where p_i is the probability of state i. The convention $0\log 0 = 0$ is used. With the natural logarithm, entropy is measured in nats; with logarithm base 2, it is measured in bits.

A sequence dominated by one state generally has lower entropy than a sequence in which several states occur with similar probabilities. In agricultural research, entropy may help compare the diversity of rainfall states, temperature categories, or soil-moisture conditions across locations and seasons.

Researchers can also examine the frequency of symbolic patterns. For instance, a sequence such as D → D → H might represent two dry observations followed by heavy rainfall. The frequency of such sequences can be compared across periods or regions. However, identifying a recurring pattern does not automatically establish its agricultural significance; interpretation requires appropriate context and validation.

Symbolic Regression for Agricultural Relationships

Symbolic regression is distinct from SDA and symbolic time series analysis. Rather than representing observations as intervals or studying transitions between symbols, symbolic regression searches for mathematical expressions that describe relationships in data (Schmidt & Lipson, 2009).

A general regression relationship can be written as

$$y = f(x_1, x_2, \ldots, x_p) + \varepsilon, \quad (14)$$

where y is the response variable, $x_1, \ldots, x_p$ are explanatory variables, f is the function to be estimated, and ε represents unexplained variation. In agriculture, y might represent crop yield, while explanatory variables could include rainfall, temperature, and soil moisture. A conventional linear regression model could be written as

$$Y_i = \beta_0 + \beta_1 R_i + \beta_2 T_i + \varepsilon_i, \quad (15)$$

where Y_i is crop yield for observation i, R_i is rainfall, T_i is temperature, and $\beta_0, \beta_1, \beta_2$ are model coefficients. Symbolic regression can search for nonlinear expressions when the relationships are more complex. For example, an illustrative candidate equation is

$$Y_i = \beta_0 + \beta_1 R_i - \beta_2 T_i^2 + \varepsilon_i. \quad (16)$$

This expression allows a curved association between temperature and yield. It is only an example: it does not establish that crop yield follows this relationship, and the coefficient signs and functional form must be estimated and tested using relevant data.

Because symbolic regression may generate many candidate equations, model selection should consider both predictive error and complexity. A general penalised objective is

$$J(f) = \frac{1}{n}\sum_{i=1}^{n}[y_i - f(\mathbf{x}_i)]^2 + \lambda C(f), \quad (17)$$

where n is the number of observations, $\mathbf{x}_i$ is the vector of predictors for observation i, $C(f)$ measures model complexity, and $\lambda \geq 0$ controls the penalty for complexity. The first term measures average squared prediction error, while the second discourages unnecessarily complicated expressions.

The specific loss function and complexity measure depend on the application. Candidate models should be compared with suitable baseline methods and evaluated on independent test data or through carefully designed cross-validation.

The overall workflow of symbolic data analysis for agricultural time-series applications is presented in Figure 1. The process begins with the collection of agricultural observations, including weather, soil, crop, and yield data. After preprocessing and organization, the observations are transformed into symbolic representations, such as intervals, distributions, or categories. These representations are then used to construct symbolic time series and extract meaningful temporal patterns. Symbolic modeling and regression are subsequently applied to investigate relationships within the data. Finally, model evaluation and interpretation support agricultural forecasting and decision-making.

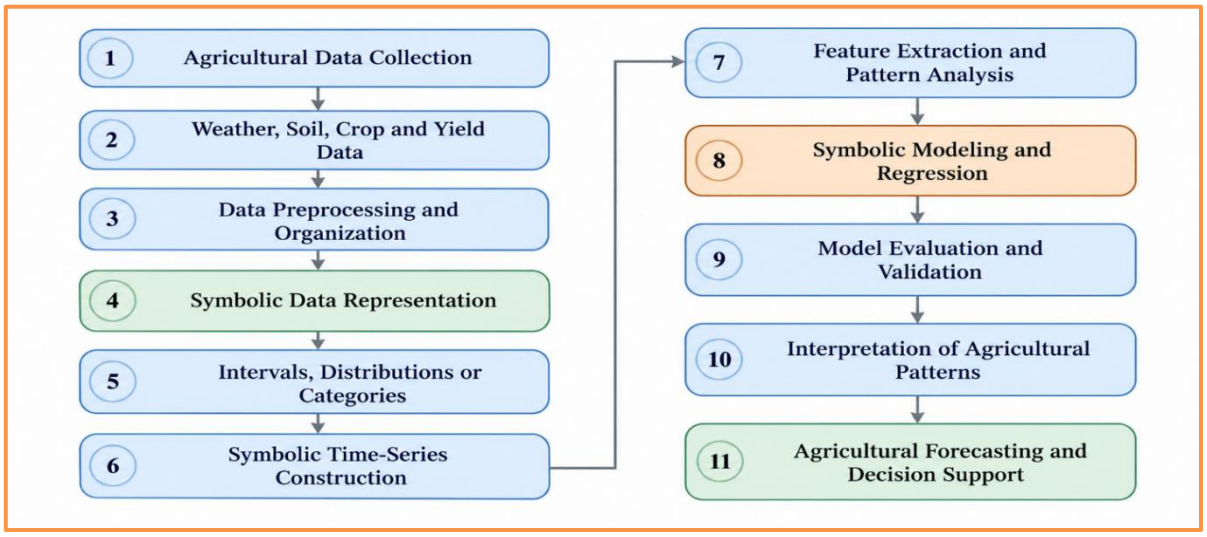


Figure 1. Workflow of Symbolic Data Analysis for Agricultural Time Series.

Potential Applications in Agricultural Research

Climate variability and extreme conditions

Interval and histogram representations can summarize the variability of temperature and rainfall across regions and growing seasons. Symbolic analysis of time series can describe transitions between dry, moderate and wet conditions or between temperature categories. These methods could help researchers to characterize climate variability and explore agricultural exposure to extreme events.

Crop growth and yield analysis

Environmental conditions can be aggregated over key crop growth stages and compared with yield records. Symbolic regression can be useful for finding interpretable relationships between rainfall, temperature and crop productivity. The resulting model needs to be evaluated on appropriate seasons, locations or independent datasets before being used to inform recommendations.

Soil moisture and irrigation management

Soil moisture observations can be represented in categories such as low, adequate and high. Transition probabilities can be used to describe the frequency of moisture transfer between categories. This information can come in addition to your current irrigation models, but not a substitute for crop specific water requirements, sensor calibration or field validation.

Comparing regions and farming systems

SDA provides tools for comparing observations that are interval-valued and distribution-valued. These tools can show regional differences that are invisible in the averages. For example, two regions with similar average rainfall may have different rainfall distributions or different frequencies of dry spells. These are potential uses of the methods and do not guaranty improved predictions or agricultural outcomes. Their practical value depends on the quality of the data, the appropriateness of the model and the results of empirical evaluation.

Challenges and Future Perspectives

The results of the symbolic analysis depend on the data representation. Important variation may be masked or spurious patterns created by poor choice of intervals or thresholds. Researchers have to justify their choices, explore alternative specifications, and ask whether their results are robust to changes in those choices. Agricultural data may also suffer from missing observations, measurement errors, seasonal effects, and spatial dependence. These features should be taken into account when preparing the data and evaluating the model. A model that works well on the data it was built on may not work so well in another season or region. It is important to differentiate the three approaches discussed in this article. SDA uses structures like intervals and histograms to represent observations; symbolic analysis of time series studies sequences of discrete states; and symbolic regression searches for mathematical expressions relating variables. The methods could be combined into an integrated workflow, but the effectiveness of that workflow should be evaluated, not assumed. Future studies could compare symbolic methods with established statistical and machine-learning approaches using weather records, remote-sensing products, soil sensors and crop-yield datasets. Such comparisons could indicate if symbolic methods have any practical benefits in interpretability, variability representation or predictive performance.

Conclusion

Agricultural systems are complex and strongly influenced by changing environmental conditions. Although averages and other conventional summaries remain valuable, they may not capture all the variation and temporal structure present in agricultural data. Symbolic Data Analysis provides ways to represent intervals and distributions. Symbolic time series analysis helps researchers investigate state transitions and recurring patterns, while symbolic regression searches for mathematical relationships between agricultural variables. These complementary approaches expand the tools available for studying climate variability, crop growth, soil moisture, and agricultural productivity. Their value lies not in replacing conventional methods in every situation, but in providing additional ways to represent and

interpret complex data. With appropriate data, clear assumptions, and rigorous validation, symbolic approaches may contribute to better understanding of agricultural systems and support evidence-based research.

References

1. Beranger, B., Lin, H., & Sisson, S. A. (2023). New models for symbolic data analysis. *Advances in Data Analysis and Classification, 17*, 659–699. https://doi.org/10.1007/s11634-022-00520-8
2. Billard, L., & Diday, E. (2006). *Symbolic data analysis: Conceptual statistics and data mining*. Wiley. https://doi.org/10.1002/9780470090183
3. Piccardi, C. (2004). On the control of chaotic systems via symbolic time series analysis. *Chaos, 14*(4), 1026–1034. https://doi.org/10.1063/1.1796071
4. Schmidt, M., & Lipson, H. (2009). Distilling free-form natural laws from experimental data. *Science, 324*(5923), 81–85. https://doi.org/10.1126/science.1165893
5. Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal, 27*(3), 379–423. https://doi.org/10.1002/j.1538-7305.1948.tb01338.x
6. Villani, C. (2009). *Optimal transport: Old and new*. Springer. https://doi.org/10.1007/978-3-540-71050-9