DATA PREPROCESSING METHODS FOR AGRICULTURAL STATISTICAL DATA: ANALYSIS AND COMPARISON

Keldibekov Jahongir, Khasanov Dilmurod

Teacher, School 4, Zafarobod Disctrict, Jizzakh, Uzbekistan
Jizzakh branch of the National University of Uzbekistan named after Mirzo Ulugbek, Jizzakh, Uzbekistan
E-mail: jahongirkeldibekov9@gmail.com, tatusf2015@gmail.com

K E Y W O R D S

Keywords: agricultural statistics, data cleaning, repeated observations, imputation, outliers, scaling, crop yield

A B S T R A C T

Agricultural statistical datasets combine measurements at different spatial and temporal scales and may contain repeated records, skewed inputs and inconsistent analysis units. This paper compares five preprocessing decisions using a public crop-yield dataset: key-based aggregation, handling of missing values, outlier treatment, feature scaling and temporal validation. The merged source contains 25,932 rows, but only 13,130 distinct country–year–crop units; 3,690 units have repeated yield values paired with different temperatures. In maize data, treating source rows as separate observations gives a temperature–yield Pearson correlation of −0.549, versus −0.481 after aggregation. Since the source has no missing cells, a controlled experiment masked 15% of test-year temperature and pesticide values over 20 random repetitions. Country-specific training medians reduced temperature reconstruction MAE from 6.09 °C (global median) to 0.63 °C. In a fixed 7-neighbor yield model evaluated on later years, min–max scaling reduced MAE from 1.780 to 0.733 t ha^{-1}. These results show that the analysis unit and scaling rule can substantially alter statistical conclusions and predictive behavior. They do not establish that one preprocessing method is universally optimal.

1. Introduction

Reliable agricultural analysis begins with deciding what each row represents. Crop yield may be recorded for a field, district, province or country, whereas a climatic attribute may describe a station, monthly period or regional average. Joining such sources can produce repeated outcomes, hidden differences in temporal resolution and misleading sample sizes. These issues matter even when every cell is filled and every number is syntactically valid. Recent work has shown that crop-yield regression is sensitive to missingness and data structure and that no single preprocessing strategy performs best under all conditions [1].

Data preparation is sometimes described as a routine sequence of deletion, imputation, normalization and model fitting. For agricultural records, the order is consequential: one must first define a valid observation unit, then separate training from evaluation periods, learn transformation parameters from training data, and finally check whether a transformation preserves the substantive meaning of the variables. A large pesticide total for a country can be a genuine national value, even if it appears extreme in a model of one crop. Similarly, a low yield may reflect drought rather than a sensor error. Field-scale yield-map cleaning has therefore emphasized robust statistics together with explicit knowledge of plausible measurements [2].

The purpose of this article is to provide an auditable comparison of preprocessing methods on open agricultural statistics. The questions are: How much does aggregation change the unit count and correlation estimates? How do simple missing-value strategies perform when values are withheld for evaluation? How do capping and transformations affect a skewed predictor? And how does scaling alter a distance-based crop-yield model under a later-year test? The contribution is an empirical comparison tied to clearly stated units and evaluation conditions, rather than an unsupported claim that cleaning automatically increases model accuracy.

(cc) BY

2. Data Source and Audit

2.1. Public dataset and measurement levels

The experiment uses the merged CSV deposited by Laktionov, Semenov and Diachenko [3]. According to the record, crop yield and pesticide totals originate from FAO sources, while rainfall and temperature were drawn from another public crop-yield collection. Its seven fields describe country, year, crop, rainfall, pesticide use, temperature and yield. We retain the original units: annual rainfall in mm, pesticide total in tonnes, temperature in °C and yield in hg ha^{-1}. For model errors, yield is converted to t ha^{-1} by division by 10,000. The file includes 101 countries, ten crop classes and years 1990–2013, with 2003 absent.

Table 1 records the data dictionary and the key interpretation used here. The numerical fields do not all have comparable scope: pesticide mass is a national total, rainfall has one fixed value for each country over the entire time series, and temperature can occur more than once for the same crop and year. This prevents rainfall from representing annual weather anomalies and prevents pesticide totals from being interpreted as per-hectare doses or applications to a particular crop.

Table 1.

.Variables and the analysis level in the deposited dataset

Field	Unit or type	Observed level / caution
area	Country name	Geographic identifier
year; item	Year; crop category	Candidate country–year–crop key
average_rain_fall_mm_per_year	mm	Constant within each country in this file
pesticides_tonnes	tonnes	National total shared across crops in a year
avg_temp	°C	Several values may map to one yield unit
value_hg_ha	hg ha^{-1}	Identical inside every repeated key group

An exact-row duplicate check alone would fail to find the most consequential repetitions, because the temperature entries within one key group differ. The declared observation unit for all outcome analyses is therefore one country–year–crop combination. When a group contains several temperature values but a single yield, the primary analysis uses their arithmetic mean; a median alternative is retained for sensitivity analysis. Because the source does not explain what each temperature repetition represents, neither estimator is claimed to recover a specific growing-season temperature.

2.2. A concrete source-row example

The example in Table 2 illustrates why key-based aggregation differs from indiscriminate deletion. Australia's maize yield in 2000 appears alongside six temperature entries in the source; its yield and other country-year attributes agree. Treating those lines as six independent maize harvests would weight Australia in 2000 six times. The aggregation keeps one yield and calculates a mean temperature from the six available values.

Table 2.

Example of repeated measurements and the retained observation

Country–year–crop	Source rows	Temperature summary	Retained yield
Albania–1990–maize	1	16.37 °C	36,613 hg ha^{-1}
Australia–2000–maize	6	16.67 °C mean	49,360 hg ha^{-1}

(cc) BY

Country–year–crop	Source rows	Temperature summary	Retained yield
Brazil–2010–maize	9	23.64 °C mean	43,667 hg ha^{-1}
India–2011–maize	19	25.82 °C mean	24,784 hg ha^{-1}

3. Preprocessing Methods and Experimental Design

3.1. A reproducible sequence

The proposed sequence is shown in Figure 1. Schema checks include missing-cell counts, numeric ranges, categorical spelling and time coverage. Key checks count group multiplicity and test whether the outcome varies inside groups. After aggregation, the data are split by year before any imputation statistics, quantile limits or scale parameters are estimated. The main predictive experiment uses maize because it is represented in many countries while avoiding yield-level differences among crop categories. We use no target value to reconstruct an input.

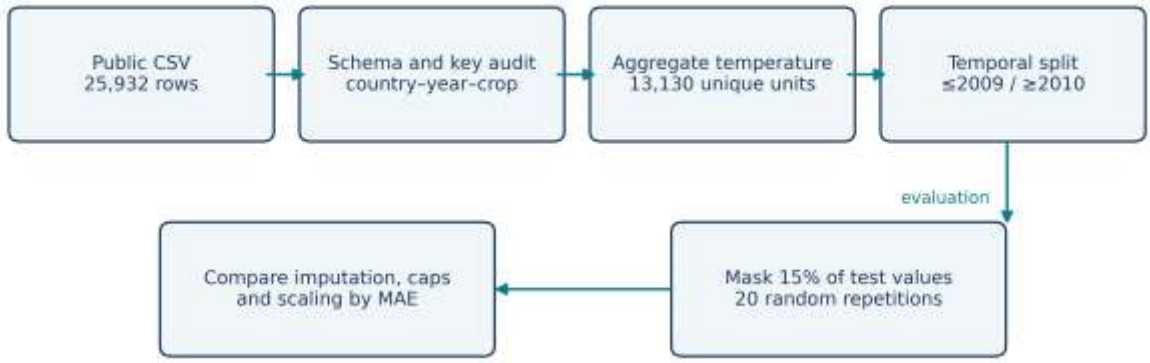


Figure 1. Audit, aggregation, temporal split and comparative evaluation

The operations have different aims. Aggregation restores the unit of inference; imputation addresses unavailable measurements; outlier handling changes the influence of extreme predictors; scaling changes numerical distances in a model. Their effects should be reported separately, since an improvement in prediction error is not proof that an observation is invalid or that an imputed measurement is biologically correct.

3.2. Key-based aggregation and correlation

Let G(c,t,k) denote source rows for country c, year t and crop k. We retain yield if all members of G have the same value, and set the group temperature to the arithmetic mean. A median temperature variant provides a robust comparison. If group yields disagree, the group would require an explicit adjudication rule instead of an automatic mean; no such conflicts occur in this file. Pearson correlations are calculated for maize between yield and each numeric attribute under three alternatives: all original rows, one row per key with mean temperature, and one row per key with median temperature.

$$T = (1 / n)\, \Sigma_j T_j\,; \quad Y = Y_j$$

The comparison examines the weighting induced by repeated records. It does not isolate a true causal temperature effect: rainfall and pesticide measures represent broader geographic or national conditions, and the data omit irrigation, soil, cultivar and planting decisions. A key-based check is more informative than a count of exact duplicates because it exposes the mismatch between predictor multiplicity and outcome multiplicity.

3.3. Missingness and imputation

The deposited CSV has zero missing cells in its seven fields. To compare imputation methods without inventing real-world gaps, a controlled missingness experiment hides 15% of observed values in the later-year maize test set, separately for temperature and pesticide use. Twenty reproducible random selections are used (seeds

(cc) BY

2026–2045). About 55 of the 366 test observations are masked for each variable in each repetition. The original values are kept only for evaluating reconstruction error. This design approximates missing completely at random (MCAR); it does not represent sensor failures or other selective missingness.

We compare the global training mean, global training median and a country-specific training median, using the global median if a test country lacks training history. All estimates use years through 2009 only. Absolute reconstruction errors are averaged across masked values, then across the twenty repetitions. Country medians can be particularly accurate when a variable changes slowly within countries, but they may fail after abrupt shifts or for countries unseen during training. The result is specific to the data and masking mechanism.

$$MAE = (1 / M)\, \Sigma_{\mathrm{i}} \in H\, |x_{\mathrm{i}} - \hat{x_{\mathrm{i}}}|$$

Here H denotes the set of M deliberately hidden test values, x is the recorded original value and x̂ is its estimate. The metric is reported in °C for temperature and tonnes for national pesticide use. Since the latter has a long right tail, we also report variability across repetitions rather than present a single run as definitive.

3.4. Outlier handling and scaling

An interquartile-range (IQR) screen identifies unusually large national pesticide totals. The conventional upper fence is Q3 + 1.5(Q3 − Q1); it flags candidates for inspection and is not a deletion rule. We compare the original maize pesticide–yield correlation with correlations after two explicitly different transformations: capping pesticide totals to training-derived 1st and 99th percentiles, and applying log(1 + pesticide total). Yield remains untouched. Capping limits influence, whereas the logarithm changes spacing throughout the positive scale; neither produces a crop-specific application rate.

$$x* = min[max(x, q_{0.01}), q_{0.99}]$$

For model inputs we compare no scaling, min–max scaling, standard scaling and robust scaling. Min–max maps the training minimum and maximum to a unit interval; standard scaling subtracts the training mean and divides by its standard deviation; robust scaling subtracts the median and divides by the IQR. Fitting a transformation on all years would leak knowledge of the later-year distribution into training. The transformations are therefore fitted within the training pipeline, as recommended in the official scikit-learn guidance [4], [5].

$$z = (x - \mu) / \sigma\,; \quad z_{\mathrm{r}} = (x - median) / IQR$$

3.5. Validation protocol

The maize panel consists of 2,028 unique country–year observations. Training years through 2009 contain 1,662 observations; 2010–2013 form a 366-observation test set. No 2003 records exist. The response is yield in t ha⁻¹. Predictors are year, the three numeric variables and a one-hot encoded country identifier. All variants use the same 7-neighbor regression model with distance weights, without hyperparameter search. The comparison is deliberately a sensitivity experiment for preprocessing, not a search for the best forecasting system.

Training preprocessing is contained in a pipeline. Unseen test countries are allowed by the categorical encoder and receive zeroes for unobserved country categories; there is no country exclusion from the training set when that country has earlier records. Reported metrics are test mean absolute error (MAE), root mean squared error (RMSE) and R^2. For the outlier experiment, the cap limits are learned from the training subset. A separate raw-row training variant retains repeated source rows but evaluates on the same 366 unique test outcomes, so changes in error reflect different training weights rather than a different test target.

$$MAE = (1 / n)\, \Sigma i\, |yi - \hat{y}i|\,; \quad RMSE = \sqrt{[(1 / n)\, \Sigma i\, (yi - \hat{y}i)^2]}$$

4. Results

4.1. Multiplicity and statistical consequences

The audit finds 25,932 source rows but only 13,130 unique country–year–crop keys. Thus 12,802 rows exceed the count of distinct outcome units; this is a multiplicity difference, not a claim

(cc) BY

that the excess rows are exact copies. In 3,690 key groups the temperature entries differ while the yield agrees. Group size reaches 19. Figure 2 shows that 9,440 units have one source row, 2,575 have two to five, 931 have six to ten and 184 have eleven to nineteen. All fields are complete, but the outcome would be unevenly weighted without this audit.

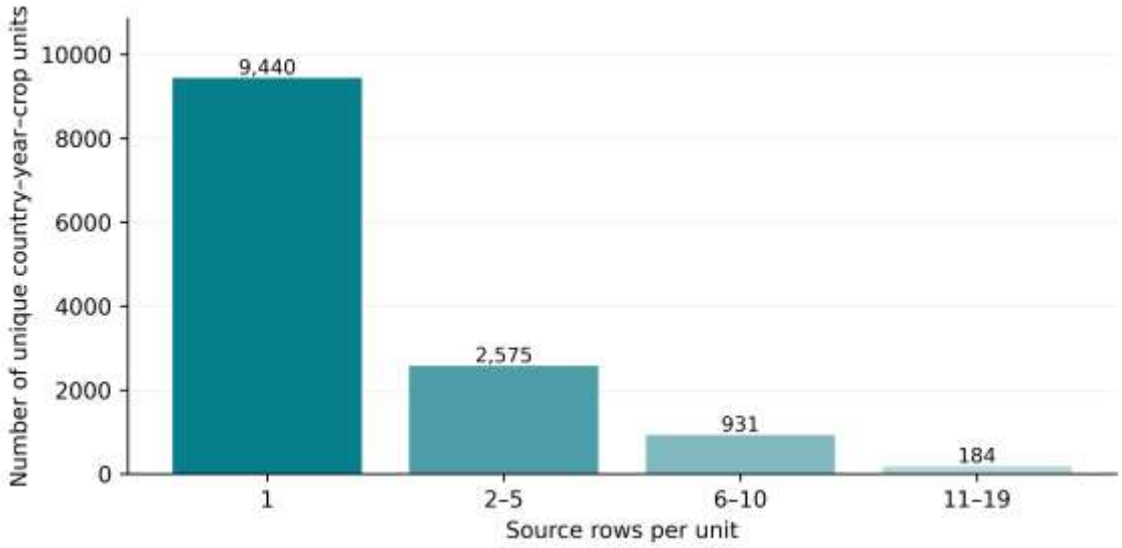


Figure 2. Source-row multiplicity for each country–year–crop unit

Table 3 and Figure 3 quantify the effect for maize. Unaggregated source rows yield a temperature correlation of −0.549 and pesticide correlation of +0.087; one observation per country–year–crop gives −0.481 and +0.194, respectively. The median temperature aggregation changes only the temperature coefficient to −0.477. Rainfall changes from −0.190 to −0.131 solely because the row weights change. The comparison demonstrates that a repeated outcome can alter a published descriptive statistic even without any missing values.

Table 3.

Pearson correlations with maize yield under alternative observation units

Preprocessing choice	Rows	Rainfall r	Temperature r	Pesticides r
Original source rows	3,824	−0.190	−0.549	+0.087
Unique units; mean temperature	2,028	−0.131	−0.481	+0.194
Unique units; median temperature	2,028	−0.131	−0.477	+0.194

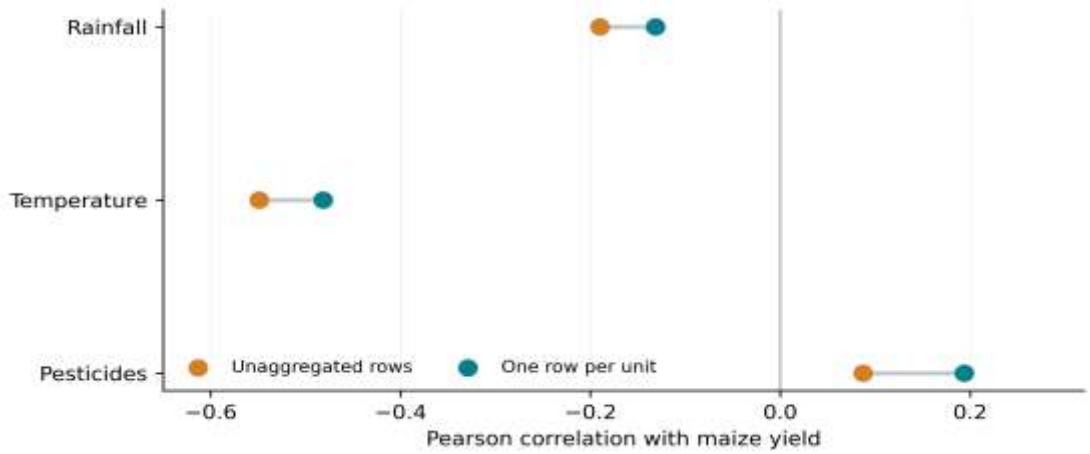


Figure 3. Change in maize correlations after key-based aggregation

(cc) BY *dtai.uz*

4.2. Controlled imputation comparison

Table 4 and Figure 4 show the mean reconstruction MAE across twenty masks, with the between-repetition standard deviation. Country-specific medians achieve 0.63 °C versus 6.09 °C for a global temperature median. For national pesticide totals, the country median also has the lowest MAE, 8,806 tonnes, but variability remains large (standard deviation 3,876 tonnes across masks). The advantage of country information is consistent with persistent country effects in this dataset; it cannot be generalized to previously unseen countries or abrupt changes in pesticide use.

Table 4.

Recovery of 15% artificially hidden later-year input values

Masked variable	Global mean MAE ± SD	Global median MAE ± SD	Country median MAE ± SD
Temperature (°C)	6.10 ± 0.34	6.09 ± 0.39	0.63 ± 0.16
Pesticides (tonnes)	21,213 ± 5,886	18,272 ± 6,586	8,806 ± 3,876

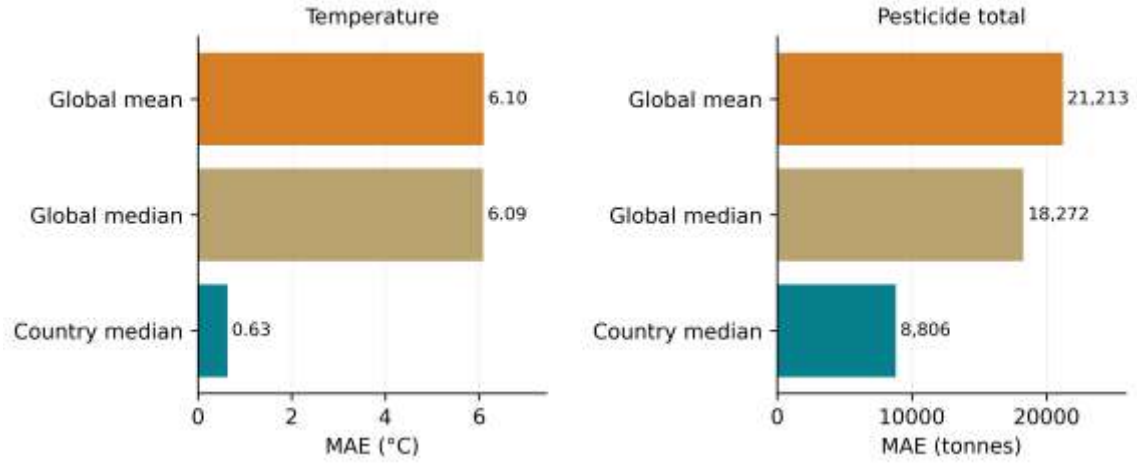


Figure 4. Mean imputation error from twenty independent masks

An important methodological distinction is that these masks were generated for evaluation only. The genuine source file remains complete, and the model comparison below uses its observed inputs. Imputation was assessed by reconstruction error rather than by selectively substituting invented gaps into the crop-yield model. Other missingness mechanisms, particularly gaps concentrated in extreme weather years, could lead to a different ranking [1].

The maize training pesticide distribution is highly skewed: Q1 is 180.4 tonnes and Q3 is 11,934.0 tonnes, giving an upper IQR fence of 29,564.4 tonnes. A total of 244 of 1,662 training observations exceed that fence. This flags large national values; it does not prove measurement error. The 99th training percentile is 104,245.27 tonnes, exceeded by 17 training records. The 1st percentile is 0.30 tonnes. These limits are used only in the controlled cap variant.

4.3. Skewed inputs and outlier decisions

Table 5.

Effects of pesticide transformations on maize correlation

Pesticide representation	Pearson r with yield	Interpretation
Original national total	+0.194	Sensitive to the raw scale and high totals
Cap at training 1st–99th percentiles	+0.255	Influence of extremes reduced

(cc) BY

Pesticide representation	Pearson r with yield	Interpretation
log(1 + national total)	+0.373	Spacing compressed across the entire range

Table 5 shows that the correlation changes when the predictor is transformed; the physical yield observations are unchanged. It would be unsound to select a transformation simply because it produces a larger coefficient. When the research question is exposure to pesticide treatment, the needed variable is a crop-specific application rate, preferably by cultivated area and season. The country total in this dataset cannot supply that interpretation.

4.4. Downstream sensitivity to scaling

With the model, split and target held constant, scaling has a substantial effect (Table 6 and Figure 5). On the 2010–2013 unique-unit test set, unscaled inputs give MAE 1.780 t ha^{-1} and R^2 0.422, while min–max scaling gives MAE 0.733 and R^2 0.875. Standard and robust scaling give MAE 0.795 and 0.770. The fixed model uses Euclidean proximity: the large numeric scale of national pesticide tonnes overwhelms year and temperature unless features are scaled. These comparisons concern a specific kNN setup and are not claims about models that are insensitive to monotone rescaling.

Table 6.

Maize yield test performance across preprocessing choices

Configuration	MAE (t ha^{-1})	RMSE (t ha^{-1})	R^2
No numeric scaling	1.780	2.594	0.422
Min–max scaling	0.733	1.204	0.875
Standard scaling	0.795	1.234	0.869
Robust scaling	0.770	1.242	0.867
Robust + 1st–99th percentile cap	0.759	1.231	0.870
Raw-row training + robust scaling	0.727	1.216	0.873

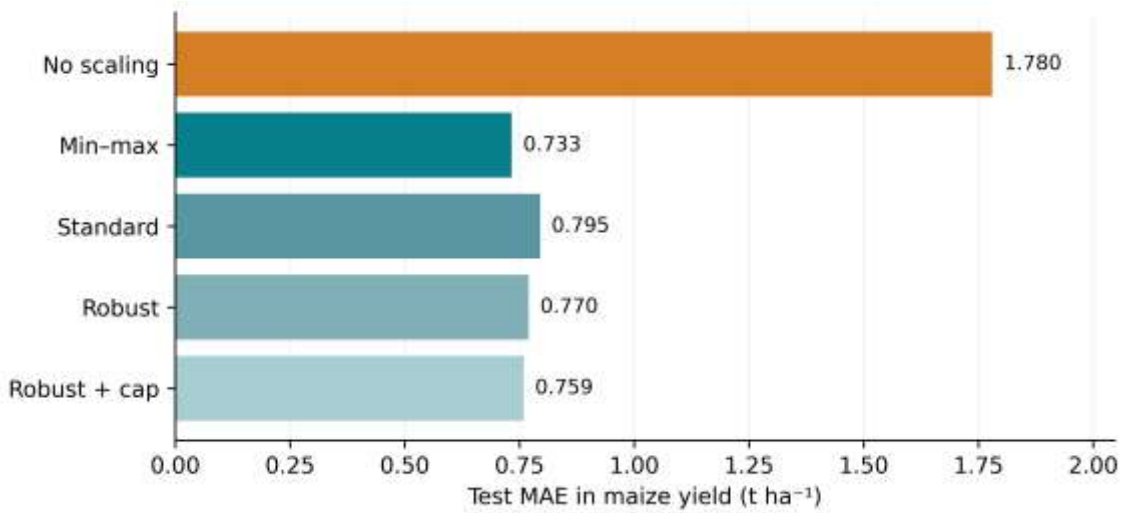


Figure 5. Test MAE for the fixed seven-neighbor model

The raw-row training variant attains MAE 0.727, slightly below the aggregated robust variant (0.770). This is a useful counterexample to the assertion that deduplication must improve every prediction metric. Repeated source lines can act as unequal training weights, sometimes helping one fixed test set by chance or by country composition. They still should not be counted as independent yield observations in inferential statistics. A preprocessing decision should therefore be

(cc) BY

justified by the data-generation process and the intended unit of analysis as well as by predictive error.

5. Discussion and Limitations

The comparison separates three questions often conflated in agricultural statistics. First, does the dataset faithfully represent the unit being analysed? Key-based aggregation is needed here because repeated temperatures accompany one crop yield. Second, how should genuine gaps be addressed? Since no gaps exist in the deposited file, our masking experiment only characterizes method behavior under randomly hidden values. Third, what numerical representation serves a particular analysis or model? Capping, logs and scaling change different properties and should not be treated as interchangeable cleaning steps.

The strongest caveat concerns provenance and measurement scope. The dataset record identifies upstream sources, but the merged file does not define the repeated temperature entries. Mean aggregation restores one row per crop outcome while leaving uncertainty about the temperature summary. Rainfall is country-constant and cannot quantify within-country annual precipitation shocks. Pesticide totals are shared across crops and do not record what was applied to maize. Soil properties, fertilizer, irrigation, planting date, crop variety and biotic stress are unobserved. None of the correlations or model outputs can isolate causal effects of temperature, rainfall or pesticide application.

The imputation experiment assumes random hiding and reuses countries with earlier years. If future data consist of new countries or if missingness is concentrated during unusual growing seasons, country medians could perform poorly. The temporal holdout tests later years but is a single split; uncertainty from alternate boundaries and model hyperparameters was not estimated. The seven-neighbor model serves as a controlled sensitivity instrument, not a production forecast. The apparent advantage of min–max scaling may change for another distance metric, crop or held-out period. These constraints align with broader evidence that preprocessing choices should respond to the dataset's structure [1].

A stronger future study would record source timestamps and spatial footprints for each climate measurement, seasonal rather than long-term rainfall, crop-specific pesticide and fertilizer rates, soil characteristics, and field identifiers. It could then compare missingness mechanisms, use group-aware or forward-chaining validation, and assess both prediction and scientific interpretation. The complete preprocessing rule, including key conflict handling and train-derived transformation parameters, should be archived with any reported result.

6. Conclusion

On the analysed public data, the first and most consequential preprocessing decision is defining one country–year–crop outcome unit. This reduces 25,932 rows to 13,130 unique units and changes maize correlations materially. Simulated missing values show that a country-specific training median can outperform a global median when country history is available; skewed pesticide totals respond differently to capping and log transformation. In a fixed later-year maize prediction task, min–max scaling lowered MAE from 1.780 to 0.733 t ha^{-1}. No single method is preferred solely by these figures: preprocessing should preserve the measurement meaning, be fitted without test leakage, and be evaluated for its stated analytical purpose.

The input is the public df_yield_merged.csv from Zenodo [3]. All numerical results reported here were recomputed from that file. The unit key, temperature aggregation, time split, random seeds, percentile definitions, fixed kNN settings and error metrics are specified in Sections 3–4. Source records remain unchanged; artificial gaps are introduced only inside the imputation evaluation.

References

1. M. L. Zinzinhédo, M. F. Mitchozounnou, K. V. Salako, and R. Glèlè Kakaï, "Sensitivity of machine learning regression models to data structure and quality in crop yield prediction," PLOS ONE, vol. 21, no. 9, e0353938, 2026, doi: 10.1371/journal.pone.0353938.
2. P. Arizo-García, S. Castiñeira-Ibáñez, E. Cruzado-Campos, B. Ricarte, C. Rubio, and A. San Bautista, "A standardized framework

(cc) BY

for cleaning non-normal yield data from wheat and barley crops, and validation using machine learning models for satellite imagery," Agronomy, vol. 16, no. 3, 386, 2026, doi: 10.3390/agronomy16030386.

3. I. Laktionov, S. Semenov, and G. Diachenko, "Dataset for Crop Yield Prediction," Zenodo, version 1, 2026, doi: 10.5281/zenodo.18761663.
4. scikit-learn developers, "Common pitfalls and recommended practices," scikit-learn User Guide, 2026. Available: https://scikit-learn.org/stable/common_pitfalls.html.
5. N. Raximov, D. Khasanov, "THE IMPORTANCE OF FORECASTING CROP YIELDS AND REGIONAL CROP DISTRIBUTION BASED ON AGRO-CLIMATIC CHARACTERISTICS AND MARKET DYNAMICS," Raqamli Transformatsiya va Sun'iy Intellekt, 4(5), 1-5, 2026. https://dtai.uz/articles/the-importance-of-forecasting-crop-yields-and-regional-crop-distribution-based-on-agro-climatic-characteristics-and-market-dynamics/
6. D. Khasanov, N. Raximov, N. Raximov, "PREDICTION THE YIELD OF GRAIN CROPS USING BASIC MACHINE LEARNING ALGORITHMS," Raqamli Transformatsiya va Sun'iy Intellekt, 2(5), 40-46, 2024. https://dtai.uz/articles/prediction-the-yield-of-grain-crops-using-basic-machine-learning-algorithms/
7. D. Khasanov, B. Daminova and M. Tojiyev, "The Impact of Normalization on Regression-Based Crop Yield Prediction: Accuracy and Efficiency Analysis," *2026 International Russian Smart Industry Conference (SmartIndustryCon)*, Sochi, Russian Federation, 2026, pp. 63-67, doi: 10.1109/SmartIndustryCon68821.2026.11493015.
8. Dilmurod, H., and R. Shirinboyev. "THE MAIN THEORY OF LINEAR REGRESSION MODELS IN DATA MINING." Sun'iy Intellekt Nazariyasi va Amaliyoti: Tajribalar, Muammolar va Istiqbollari (2024): 143-147.
9. Shirinboyev, R., Rakhimov, N., Khasanov, D., & Xafizadinov, U. (2024). EVALUATING THE PERFORMANCE OF CONVOLUTIONAL NEURAL NETWORKS AND HYBRID CNN-SVM MODELS FOR SYMBOL RECOGNITION IN COMPLEX DATASETS. DTAI–2024, 1, 295-299.
10. Nodir, Rakhimov, Khasanov Dilmurod, Turakulov Olim, Parmonov Abdutolib, and Dusanov Xurshid. "Development of an algorithm for selecting repeat crops type and predicting yield in the conditions of agriculture in Uzbekistan." Plant Science Today 13, no. 3 (2026).

(cc) BY *dtai.uz*