Mo’ Data Mo’ Snow Imputation Standardization How-To Manual

Data imputation remains an underdeveloped and often misunderstood area in data science.Many practitioners rely on conventional techniques like mean imputation, linear interpolation, forward or backward fill, or dropping NaNs. However, these approaches may be ill-suited for datasets with high missingness, non-linear behavior, or strong seasonal patterns. Using an inappropriate imputation method can introduce bias, distort trends, and produce unreliable results. Imputation for time series demands careful attention to preprocessing, method selection, and validation. Robust analysis must include masking strategies that simulate real-world scenarios, proper transformation and stationarity testing, and performance evaluation over multiple runs. To select the most suitable imputation approach for your dataset, consider the following critical factors:

- Data gap size,
- Seasonality,
- Data patterns,
- Data formats,
- Suitability of the method to your domain
- Validation metrics (MAE, RMSE)
- Missing data mechanism (MCAR, MAR, MNAR)

In the preprocessing phase, there are a series of considerations that should be accounted for:

1. **Standardize Formats and Structures:** To ensure high-quality imputation, several preprocessing steps are crucial. First, time series data should be standardized in terms of formats and structures. This includes aligning all datasets to a common datetime index and, for geospatial data, using consistent spatial formats such as GeoDataFrame in Python or `sf` objects in R. Merging datasets with different temporal coverage should be preceded by defining a target window that reflects meaningful temporal boundaries.

2. **Preserve Real Zeros:** Real zeros, such as zero precipitation or snow accumulation, must be preserved throughout processing, which can be done using masking techniques in both Python and R. These real zeros must be protected during transformations like log-scaling or differencing, which are necessary when preparing data for models that assume stationarity. If datasets span different date ranges, define an appropriate target window that captures the most meaningful period. After aligning them, address missing values via imputation, guided by method testing. In Python, apply zero-masking decorators during both testing and imputation. In R, use logical masks and reapply zeros post-imputation. Some methods and packages (e.g., `MICE`, `Amelia`, `imputeTS`) offer support for differentiating between zeros and NAs.

3. **Handle Duplicates and Conflicts:** If duplicate or overlapping columns exist across datasets, rename or suffix the columns to distinguish them. This issue can arise when merging based on spatial criteria (e.g., within 40 km of a basin). It is important to exercise caution under these conditions when even one column is dropped. If not properly treated, it can lead to unresolved

duplication, imputation bias, feature redundancy, and metadata misalignment. To mitigate these risks use suffixes (e.g., `_x, _y)` to retain both versions temporarily; compare values across columns (e.g., `df[col_x] == df[col_y]`) to check for true duplication; and only aggregate after reviewing distributions (e.g., via `.describe()` or histogram plots).

For duplicate rows, iIf both timestamp and values are identical, drop one. IIf timestamps match but values differ, aggregate using mean, median, or another contextually appropriate method.

4. **Assess level of missingness:** Common practice drops columns with >50–60% missing data. However, a threshold test can help set this cutoff more optimally. With the later strategy, it is important to exercise caution so as not to discard significant amounts of data that will make complex imputation infeasible due to low coverage. Additionally, Selecting an imputation method should be based on a procedural understanding of data features and the underlying missing data mechanism (MCAR, MAR, or MNAR). For instance, mean imputation is only appropriate under MCAR with small gaps, whereas methods like Kalman filtering or Random Forests are better suited for MAR or complex patterns. Hybrid models, for example, combine temporal linear interpolation for short gaps and Random Forests for longer or seasonal gaps. They can yield strong performance by balancing precision and pattern recognition. However, caution must be exercised with techniques like Kalman smoothing, which can overfit and smooth out true variability if the model assumptions are violated.

5. **Preparation of dataset for timeseries:** Stationarity is a cornerstone for many time series models. The Augmented Dickey-Fuller (ADF) test is typically used to assess linear stationarity, but in cases where nonlinearity is suspected, especially in environmental or economic data, the Kapetanios–Shin–Snell (KSS) test offers a more sensitive alternative. The ADF test evaluates the null hypothesis that a time series is nonstationary, whereas the KSS test checks for nonlinear mean reversion. If either test indicates nonstationarity, differencing or decomposition methods such as STL should be applied. Differencing removes stochastic trends, while regression or decomposition is suitable for deterministic trends and seasonal structures. The ADF should be conducted before imputation testing begins and post-imputation. Once non-stationary data has been identified, the data must be transformed. Zeros that are real observations should be masked, particularly since logs cannot be performed on zeros. The ADF is available in R and Python **adfuller()** (statsmodels), **tseries.**

6. **Structuring training and testing:** Once preprocessing is complete, imputation testing must be conducted using validation strategies that reflect realistic missingness. A random 20–80% train-test split is inappropriate for time series because it breaks temporal continuity. Instead, imputation methods should be validated using a mix of structured (block) and unstructured (random pointwise) masking. Block missingness mimics real-world outages or seasonal gaps, while random masking evaluates the method's robustness across scattered gaps. Importantly, evaluations should be averaged over multiple seeds to avoid overfitting to a specific missingness pattern and to ensure reproducibility. Metrics such as Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), and stationarity-preserving checks should be used to assess performance. Post-imputation, stationarity tests like the ADF must be rerun to confirm that the imputation did not falsely introduce or obscure stationarity.

Imputation Approaches for Different Datasets and File Types

Imputation Method	Data Gap Size	Patterns	Timeseries Compatibility	Data Formats	Suitability & Rationale	Validation Strategy	MCAR/ MAR/ MNAR Suitability	Computational Intenstity	References
Mean Imputation	Small gaps	Simple, predictable data patterns	No	CSV, DBF, NetCDF	Simple and quick; tends to underestimate variance and distort data; use as baseline or initial benchmark only	Visual inspection, RMSE, MAE	MCAR: Yes MAR: No MNAR: No	Low	Little & Rubin, 2002
Linear Interpolation	Small gaps (days)	Short gaps, predictable changes	Yes	CSV, DBF, Parquet	Simple, efficient for smooth, small gaps; preserves continuity	Visual inspection, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	Low	Moritz & Bartz-Beielstein, 2017
Spline Interpolation	Small Moderate (days-weeks)	Non-linear seasonal patterns	Yes	CSV, DBF	Captures non-linear trends, avoids abrupt changes	Visual, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	Medium	Moritz & Bartz-Beielstein, 2017
Kalman Filtering	Moderate/ Large - weeks-months	Strong seasonal cycles/trends	Yes	CSV, Parquet, Shapefiles	Robust, handles trends and autocorrelation effectively	Cross-validation, statistical checks (RMSE, MAE)	MCAR: Yes MAR: Yes MNAR: No	Medium	Harvey, 1990; Moritz & Bartz-Beielstein, 2017
ARIMA	Moderate (weeks)	Seasonal patterns, requires stationarity	Yes	CSV, Parquet	Good for moderate seasonal cycles; stationarity needed	RMSE, MAE, visual checks	MCAR: Yes MAR: Conditional MNAR: No	Medium	Box & Jenkins, 1970
ARIMA Hybrid	Moderate/ Large	Complex seasonal and trend patterns	Yes	CSV, Parquet	Improved accuracy over simple ARIMA models for complex data	Cross-validation, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	High	Zhang, 2003

Imputation Method	Data Gap Size	Patterns	Timeseries Compatibility	Data Formats	Suitability & Rationale	Validation Strategy	MCAR/ MAR/ MNAR Suitability	Computational Intenstity	References
Spatial Kriging (Geostatistical Interpolation)	Moderate to Large	Spatial correlation, smooth spatial gradients	No	NetCDF, Shapefiles, GeoTIFF	Strong geostatistical approach using spatial autocorrelation for interpolation; best with dense spatial data	Cross-validation, spatial leave-one-out, RMSE, MAE	MCAR: Yes, MAR: Yes, MNAR: No	High	Cressie, 1993
Graph-Based Imputation	Moderate to Large	Structured, relational data	No	NetCDF, JSON, GraphML, HDF5	Leverages graph structure to infer missing values, suitable for sensor or networked data	Cross-validation, graph-based error propagation metrics	MCAR: Yes, MAR: Yes, MNAR: Partial	High	Kipf & Welling, 2016
Bayesian Spatial Models	Moderate to Large	Spatial correlation, uncertainty modeling	No	NetCDF, GeoTIFF, Shapefiles	Incorporates prior knowledge and quantifies uncertainty in spatial context	Incorporates field expertise and quantifies uncertainty in spatial contexts	Posterior predictive checks, RMSE, MAE	Very High	Banerjee et al., 2004
Spatiotemporal Cubic Splines	Moderate	Continuous spatiotemporal trends	Yes	NetCDF, CSV, GeoTIFF	Good for smoothly varying spatiotemporal data; maintains continuity and curvature	RMSE, visual inspection, residual diagnostics	MCAR: Yes, MAR: Yes, MNAR: No	Medium	Wood, 2017
Tensor Completion	Large	Multi-dimensional missing patterns (e.g., time × space × variable)	Yes	NetCDF, HDF5, Parquet	Effective for high-dimensional structured datasets like remote sensing	RMSE, MAE, reconstruction error	MCAR: Yes, MAR: Yes, MNAR: No	Very High	Liu et al., 2013

Imputation Method	Data Gap Size	Patterns	Timeseries Compatibility	Data Formats	Suitability & Rationale	Validation Strategy	MCAR/ MAR/ MNAR Suitability	Computational Intenstity	References
Inverse Distance Weighting (IDW)	Small to Moderate	Smooth spatial gradients	No	NetCDF, Shapefiles, CSV	Simple, efficient interpolation based on proximity; sensitive to spatial density	Leave-one-out validation, MAE, RMSE	MCAR: Yes, MAR: Yes, MNAR: No	Low	Shepard, 1968
LSTM	Moderate/ Large	Strong seasonal patterns, large gaps	Yes	CSV, Parquet, Shapefiles	Excellent for extended gaps, robust in complex temporal data	Cross-validation, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: Partial	High	Hochreiter & Schmidhuber, 1997
LTSSR	Moderate /Large	Structured seasonal data	Yes	CSV, Parquet, Shapefiles	Good for structured data with clear seasonal signals	Statistical validation	MCAR: Yes MAR: Yes MNAR: No	Medium	Silva et al., 2021
Gaussian Process Regression	Moderate/ Large	Complex data patterns	Yes	CSV, Parquet, Shapefiles	Accurate with structured seasonal data, computationally intensive	RMSE, MAE, cross-validation	MCAR: Yes MAR: Yes MNAR: No	High	Rasmussen & Williams, 2006
Gradient Boost	Moderate/ Large	Complex nonlinear relationships	Conditional	CSV, Parquet, Shapefiles	Accurate predictions, careful to avoid overfitting	Cross-validation, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	High	Friedman, 2001
Random Forest	Moderate/ Large	Complex interactions, multiple predictors	Conditional	CSV, Parquet	High accuracy and robustness to data complexity	Cross-validation, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	High	Breiman, 2001
KNN	Small/ Moderate	Local spatial -temporal patterns	No	CSV, DBF, Shapefiles	Good for local patterns; requires correlated variables	RMSE, MAE, visual checks	MCAR: Yes MAR: Yes MNAR: No	Medium	Batista & Monard, 2002

Imputation Method	Data Gap Size	Patterns	Timeseries Compatibility	Data Formats	Suitability & Rationale	Validation Strategy	MCAR/ MAR/ MNAR Suitability	Computational Intenstity	References
SVD (Truncated)	Moderate	Dimensionality reduction, structured gaps	No	CSV, Parquet	Good for dimensionality reduction; sensitive to data complexity	RMSE, MAE, visual checks	MCAR: Yes MAR: Yes MNAR: No	Medium	Halko et al., 2009
Linear Regression (LR)	Moderate	Linear relationships	No	CSV, DBF	Simple, effective with strong linear relationships	RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	Low	Kutner et al., 2004
Multiple Imputation (MICE)	Moderate	Research contexts, uncertainty quantification	Conditional	CSV, DBF	Robust statistical conclusions; computationally intensive	Cross-validation, statistical comparison	MCAR: Yes MAR: Yes MNAR: No	High	van Buuren & Groothuis-Oudshoorn, 2011; Schafer & Graham, 2002
Forward and Backward Fill	Very small gaps	Short continuity needed	No	CSV	Quick, minimalistic; avoid if significant variability	Visual check	MCAR: Yes MAR: No MNAR: No	Low	Little & Rubin, 2002
Dropping NaNs	Very Small, Sporadic	Breaks continuity	No	CSV, DBF, Shapefiles	Simple omission; valid only under MCAR and when missingness is minimal	MCAR test, deletion comparison	MCAR: Yes MAR: No MNAR: No	Low	Little & Rubin, 2002
Matrix Completion (SVD)	Moderate to Large	Latent seasonal patterns, low-rank structure	Conditional	CSV, Parquet, NetCDF, Files	Reconstructs missing values assuming low-rank approximation holds	RMSE, MAE, reconstruction diagnostics	MCAR: Yes MAR: Yes MNAR: No		Halko et al., 2009

Imputation Method	Data Gap Size	Patterns	Timeseries Compatibility	Data Formats	Suitability & Rationale	Validation Strategy	MCAR/ MAR/ MNAR Suitability	Computational Intenstity	References
Hybrid: Temporal + Random Forest	Small to Large	Complex trends, interactions, temporal cycles	Yes	CSV, Parquet, Shapefiles	Short-gap interpolation plus RF for complex recovery	Cross-validation, RMSE, MAE	MCAR: Yes MAR: Yes MNAR: No	High	Breiman, 2001
Hybrid: Temporal + KNN	Small to Moderate	Repeated loca seasonal patterns	Conditional	CSV, DBF, Shapefiles, NetCDF	Temporal interpolation + KNN for contextual similarity	RMSE, MAE, visual checks	MCAR: Yes MAR: Yes MNAR: No	Medium	Batista & Monard, 2002

References

Banerjee, S., Carlin, B. P., & Gelfand, A. E. (2004). *Hierarchical modeling and analysis for spatial data*. Chapman and Hall/CRC.

Batista, G. E. A. P. A., & Monard, M. C. (2002). A study of K-nearest neighbour as an imputation method. *His*, 87(251–260), 48.

Box, G. E. P., & Jenkins, G. M. (1970). *Time series analysis: Forecasting and control*. Holden-Day.

Breiman, L. (2001). Random forests. *Machine Learning, 45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... & Li, Y. (2017). XGBoost: Extreme Gradient Boosting. *R package version 0.6-4*.

Cressie, N. A. C. (1993). *Statistics for spatial data* (rev. ed.). Wiley.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics, 29*(5), 1189–1232.

Halko, N., Martinsson, P. G., & Tropp, J. A. (2009). Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *arXiv preprint arXiv:0909.4061*.

Harvey, A. C. (1990). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation, 9*(8), 1735–1780.

Kipf, T. N., & Welling, M. (2016). Semi-supervised classification with graph convolutional networks. *arXiv preprint arXiv:1609.02907*.

Kutner, M. H., Nachtsheim, C. J., Neter, J., & Li, W. (2004). *Applied linear statistical models* (5th ed.). McGraw-Hill/Irwin.

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data* (2nd ed.). Wiley.

Liu, J., Musialski, P., Wonka, P., & Ye, J. (2013). Tensor completion for estimating missing values in visual data. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 35*(1), 208–220.

Moritz, S., & Bartz-Beielstein, T. (2017). imputeTS: Time series missing value imputation in R. *The R Journal, 9*(1), 207–218. https://doi.org/10.32614/RJ-2017-009

Parmezan, A. R. S., & Batista, G. E. A. P. A. (2015). A study of the use of complexity measures in the similarity search process adopted by kNN algorithm for time series prediction. In *2015 IEEE 14th International Conference on Machine Learning and Applications (ICMLA)* (pp. 45–51). IEEE. https://doi.org/10.1109/ICMLA.2015.217

Rasmussen, C. E., & Williams, C. K. I. (2006). *Gaussian processes for machine learning*. MIT Press.

Schafer, J. L., & Graham, J. W. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), 147–177. https://doi.org/10.1037/1082-989X.7.2.147

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In *Proceedings of the 1968 ACM National Conference* (pp. 517–524).

Silva, D. F., Batista, G. E. A. P. A., Keogh, E., & Papa, J. P. (2021). An efficient and accurate method for time series classification and clustering using symbolic representations. *Information Sciences, 509*, 294–312.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. *Journal of Statistical Software, 45*(3), 1–67.

Wood, S. N. (2017). *Generalized additive models: An introduction with R* (2nd ed.). CRC Press.

Zhang, G. P. (2003). Time series forecasting using a hybrid ARIMA and neural network model. *Neurocomputing, 50*, 159–175.