Zero : Jurnal Sains, Matematika, dan Terapan
E-ISSN : 2580-5754; P-ISSN : 2580-569X
Volume 10, Number 2, 2026
DOI: 10.30829/zero.v10i2.30674
Page: 858-870



Jackknife and Jackknife+ Prediction Intervals for Machine Learning: Simulation and Susenas 2023 Household Expenditure

[1]Bayu Paramita
Department of Statistics and Data Science, IPB University, Bogor 16680, Indonesia

[2]Bagus Sartono
Department of Statistics and Data Science, IPB University, Bogor 16680, Indonesia

[3]Yenni Angraini
Department of Statistics and Data Science, IPB University, Bogor 16680, Indonesia

Article Info

Article history:

Accepted : 10 October 2026

Keywords:

Adaptive Kernel;
Conformal prediction;
Gaussian copula;
Jackknife+;
Prediction interval;
Susenas.

ABSTRACT

This study evaluated and compared jackknife and jackknife+ conformal prediction intervals for three machine learning regression models, namely random forest, gradient boosting, and support vector regression. We used two settings: a simulation that varied the predictor-to-sample-size ratio from 0.2 to 2.0 across ten scenarios, each with 100 replications, using synthetic data drawn from a semiparametric Gaussian copula, and an application to Susenas 2023 household expenditure for Bogor City. We assessed six method-model combinations at 90% nominal coverage using coverage rate, interval width, Winkler score, computation time, and peak memory. In the simulation, median coverage reached 91-92%, but the first quartile fell to 86-88%, and interval width was governed mainly by the base model, not the interval method. On 815 households, coverage ranged from 90.80% to 92.64%, with random forest and jackknife+ closest to nominal. Coverage failures concentrated in the right tail, so these dataset-specific findings require cautious generalization.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Bayu Paramita,
Department of Statistics and Data Science,
IPB University, Bogor
Email: bayu.paramita@apps.ipb.ac.id

1. INTRODUCTION

Over the past two decades, machine learning has expanded predictive modeling. Models such as random forest [1], gradient boosting [2], and support vector regression [3], [4] capture complex nonlinear patterns more effectively than classical statistical methods. Their output, however, is normally a point prediction with no information about uncertainty, although such information is equally important in data-driven decision making. A prediction interval quantifies predictive uncertainty by providing a range expected to contain the response at a confidence level. In linear regression and other classical methods, such intervals follow analytically from distributional assumptions. Machine learning models are generally nonparametric and nonlinear, so they require alternative approaches [5], [6].

Conformal prediction addresses this need by providing valid marginal coverage guarantees without distributional assumptions [7], [8]. It has been extended to split-conformal inference [9], [10], conformalized

quantile regression [11], and energy forecasting [12], with reviews documenting rapid uptake across applied domains [13], [14]. Two methods are the jackknife and jackknife+ [15]: the jackknife adapts leave-one-out procedures to construct prediction intervals, whereas jackknife+ additionally uses leave-one-out predictions at the test point.

Barber et al. [15] proved that jackknife+ guarantees marginal coverage of at least 1 - 2α for exchangeable data even when the algorithm is unstable. In contrast, ordinary jackknife has no such guarantee and can fail under instability. In practice, the difference depends on the base learner and dimensionality, namely predictors relative to sample size [16], [17]. Comparative evidence across base learners and dimensionality regimes remains limited [18], especially studies jointly reporting interval quality and computational cost.

A second gap concerns Indonesia. Machine learning has been applied to welfare and poverty data for small area estimation, poverty mapping, and program targeting[19], [20], [21], [22], but this work has generally stopped at point estimation. Predicted values are reported without assessing how reliably they bracket the truth, while coverage, width, and computational cost of prediction intervals have not been systematically evaluated on Susenas data. The missing element in Indonesian work is interval-based assessment, not a new algorithm.

This gap matters for policy. Household expenditure figures for Bogor City inform regional development planning and social protection targeting [23]. A point estimate alone does not indicate reliability, whereas an interval allows planners to distinguish dependable estimates from those requiring caution. A household predicted near an eligibility threshold with a narrow interval, and one with the same prediction but an interval spanning several million rupiah may require different administrative treatment, a distinction visible only through interval-based analysis.

This study's contribution is applied rather than methodological: to the authors' knowledge, it is the first evaluation of jackknife and jackknife+ prediction intervals for household expenditure estimation using Indonesian Susenas data. Three features distinguish it: both methods are evaluated across three base learners on interval quality and computational cost; a controlled simulation based on a semiparametric Gaussian copula reproduces empirical marginal distributions and rank dependence, allowing the p/n ratio to vary without abandoning realism; and intervals are examined observation by observation rather than only through aggregate metrics, revealing failure modes hidden by summary statistics.

The objectives were threefold: to characterize prediction intervals produced by jackknife and jackknife+ for random forest, gradient boosting, and support vector regression across p/n ratios; to quantify differences in coverage, width, Winkler score, computation time, and memory usage; and to identify the interval-method and base-model combination that performs best for household expenditure data in Bogor City based on Susenas 2023.

2. RESEARCH METHOD

This study used a quantitative computational design with two complementary components that shared the same base models, interval methods, and evaluation metrics. A simulation component assessed the stability of the two interval methods as the ratio of predictors to training sample size varied; an empirical component assessed their behavior on actual survey data. The design was chosen because the quantities of interest, finite-sample coverage and interval width, cannot be derived analytically for nonparametric learners and must be evaluated by repeated resampling.

2.1 Data Source and Variables

The data were the 2023 Susenas records for Bogor City obtained from Statistics Indonesia [24]. Susenas is a nationwide household survey measuring the socioeconomic conditions of Indonesian society, and the household consumption expenditure derived from it is widely used as a welfare proxy [25], [26]. The extract covered 817 households. The response variable (EXPEND) was average monthly household expenditure in rupiah, and the 67 explanatory variables came from eight questionnaire blocks (Table 1).

Table 1. Grouping of explanatory variables by Susenas questionnaire block

Block	Description	Numeric	Categorical	Total
I	Location information	0	1	1
III	Summary	4	0	4
XVII	Access to food	0	8	8
XVIII	Housing information	1	10	11
XIX	Access to financial services	0	10	10
XX	Ownership of goods	0	13	13
XXI	Household income information	0	1	1
XXII	Social protection information	0	19	19
Total		5	62	67

Note: 67 explanatory variables in total, comprising 5 numeric and 62 categorical variables.

2.2 Survey Design, Weights, and Missing Data

Susenas is a complex survey with stratification and multistage cluster selection, and its published estimates use sampling weights [24], [27]. The analysis reported here was unweighted and did not incorporate strata or primary sampling unit identifiers because the aim was to compare the finite-sample behavior of two interval construction procedures on a fixed set of records, not to produce design-based expenditure estimates for the population of Bogor City. All six method-model combinations were evaluated on the same records, so the comparison between them was unaffected by the omission; what the omission affects is the population to which the coverage figures refer (Section 3.9).

We examined item non-response block by block. The food access block (R1701-R1708) contained no "do not know" or "refuse to answer" codes. Variable R1812 carried code 8 ("do not know") for two households. Because non-response codes are not analytic categories, we removed those two records by listwise deletion, leaving 815 households (99.8% of the original extract). Given the negligible proportion removed, listwise deletion was adequate and the associated bias risk minimal [28]. No imputation was performed.

2.3 Data Preparation and Encoding

Every categorical variable was validated against the standard code domain in the Susenas metadata. Only R1803 (proof of land ownership) contained a value outside its documented domain, namely code 0, which occurred in 250 households (30.6%). Cross-tabulation with dwelling ownership status (R1802) showed that code 0 arose exactly when R1802 equaled 2 (rented) or 3 (rent-free), that is, when the question did not apply; when R1802 equaled 1 (owner-occupied), R1803 always took a value in 1-6, with no exceptions. Code 0 therefore reflected a structural skip rather than ordinary missingness and was retained as an explicit "not applicable" category, so that the information was preserved and not confounded with missing data [28], [29].

Categorical variables were then encoded according to their nature. The 56 binary yes/no variables were coded as 0/1 indicators; the urban-rural classification (R105) as a single dummy; and five nominal variables with more than two categories (R1802, R1803, R1816, R2101A, and R2202) were one-hot encoded, since they have no meaningful ordering and ordinal coding would impose spurious distances between categories [30]. R1812 ceased to be multi-category once the two non-response records were removed and was coded as a binary indicator. Encoding produced 20 dummy columns, giving a final design matrix of 815 x 82 (Table 2).

Table 2. Summary of explanatory variable encoding

Variable type	Encoding method	Source variables	Resulting columns
Numeric	None (standardised for SVR only)	5	5
Binary categorical (yes/no)	0/1 indicator	56	56
Binary nominal (R105)	Dummy variable	1	1
Multi-category nominal	One-hot encoding	5	20
Total		67	82

Three properties of this encoding bear on how the results should be read. First, all encodings were deterministic functions of the questionnaire metadata and none used the response variable, so target leakage could not occur; target encoding and learned embeddings were considered but rejected for that reason, since they require the response inside the leave-one-out loop. Second, the cardinality of the one-hot variables was low, contributing 20 columns, so the rise in dimensionality was modest (67 to 82). Third, several encoded columns were near-zero-variance: the credit access block (R1901A-R1901J) and most of the social assistance block recorded affirmative answers in under 5% of households, and only two observations populated one nominal category. Rare levels were not collapsed, since merging them would alter the substantive meaning of the questionnaire categories. Because the column count was held fixed while the training sample shrank in the simulation, sparsity increased mechanically with the p/n ratio (Section 3.9).

Exploratory screening identified two zero-variance columns, R105 (all records urban, as expected for a city) and R1813D (no household reported foamy water), and one exact duplicate, R2001I (boat ownership) identical to R2001L (flat-screen television ownership) across all records, with an implausible 42.6% ownership rate for an inland urban area; all three were flagged as screening candidates at the modeling stage. Numeric variables were standardized to zero mean and unit variance for support vector regression, for which feature scaling is standard practice [30], and the response log-transformed for that model to stabilize variance. Tree-based models are invariant to monotone transformations of the predictors and were fitted on the untransformed scale.

2.4 Machine Learning Regression Models

Random forest is an ensemble that builds many decision trees by bagging combined with random feature selection at each split, and aggregates their predictions by averaging [1]:

$$\hat{\mu}_{\mathrm{RF}}(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x) \quad (1)$$

where B is the number of trees and $T_b(x)$ the prediction of the b-th tree.

Gradient boosting builds the model additively and sequentially, each weak learner correcting the combined model built so far [2]:

$$F_m(x) = F_{m-1}(x) + \nu h_m(x), \quad 0 < \nu \leq 1 \tag{2}$$

with h_m fitted to the pseudo-residuals

$$r_{im} = -\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\Big|_{F=F_{m-1}} \tag{3}$$

where L is the loss function and ν the learning rate. The sequential structure reduces bias and is generally accurate on tabular data, but it is prone to overfitting and sensitive to hyperparameter choice [5].
Support vector regression uses the ε-insensitive loss and the principle of structural risk minimization [4], [31]:

$$L_\epsilon\big(y, f(x)\big) = \max(0, |y - f(x)| - \epsilon) \tag{4}$$

Through a kernel, the model represents nonlinear relationships by an implicit mapping to a higher-dimensional space, so that the prediction takes the form [32]

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \tag{5}$$

where α_i and α_i^* are Lagrange multipliers, K the kernel function, and b the intercept.

Hyperparameters were fixed a priori at the values in Table 3 rather than tuned within each leave-one-out fold. This was a deliberate design decision: the object of comparison was the interval construction procedure, and re-tuning inside every one of the n leave-one-out refits would both multiply the computational cost by the size of the search grid and confound the comparison, since differences between methods could then reflect differences in the tuning outcome rather than in the interval construction. The values were selected in preliminary runs on the empirical training data using five-fold cross-validation over a coarse grid (random forest: 50, 100, 200 trees; gradient boosting: 50, 100 iterations with learning rate 0.05 and 0.10; support vector regression: RBF kernel with C in {1, 10, 100} and ε in {0.05, 0.15, 0.30}), choosing the smallest configuration whose cross-validated RMSE was within one standard error of the best.

Table 3. Regression models and hyperparameter settings

Model	Hyperparameters
Random forest	Number of trees = 50
Gradient boosting	Number of iterations = 50; learning rate = 0.05
Support vector regression	RBF kernel; C = 100; ε = 0.15; predictors standardised; response log transformed

2.5 Jackknife and Jackknife+ Prediction Intervals

Conformal prediction delivers distribution-free intervals that are valid in finite samples, requiring only that the data be exchangeable [7], [8]. At nominal level $1 - \alpha$,

$$P\big(Y_{n+1} \in C^{n,\alpha}(X_{n+1})\big) \geq 1 - \alpha \tag{6}$$

The jackknife method traces back to Quenouille [33] and Tukey [34], was reviewed by Miller [35], and was connected to the bootstrap by Efron [36]. For each i in $\{1, \ldots, n\}$ the model was refitted without observation i to obtain $\hat{\mu}$-i, and the leave-one-out residual computed as

$$R_i^{LOO} = |Y_i - \widehat{\mu_{-i}}(X_i)| \tag{7}$$

Writing $q_{n,\alpha}^{+}$ for the k-th smallest element of $\{R_i^{LOO}\}$ with $k = \lceil (1-\alpha)(n+1) \rceil$, and $\hat{\mu}$ for the model fitted on all training data, the interval for a test point is

$$C_{\text{jackknife}}^{n,\alpha}\big(X_{n+j}\big) = \hat{\mu}\big(X_{n+j}\big) \pm q_{n,\alpha}^{+}\{R_i^{LOO}\} \tag{8}$$

This construction stores only the residuals, which makes it economical in memory. Barber et al. [15] showed, however, that its coverage guarantee can fail when the algorithm is unstable, which motivates the jackknife+. The jackknife+ replaces the single centre $\hat{\mu}\big(X_{n+j}\big)$ with the ensemble of leave-one-out predictions at the test point:

$$C_{\text{jackknife+}}^{n,\alpha}\big(X_{n+j}\big) = \big[q_{n,\alpha}^{-}\{\widehat{\mu_{-i}}\big(X_{n+j}\big) - R_i^{\text{LOO}}\},\ q_{n,\alpha}^{+}\{\widehat{\mu_{-i}}\big(X_{n+j}\big) + R_i^{\text{LOO}}\}\big] \tag{9}$$

where the lower endpoint is the k^--th smallest value with $k^- = \lfloor \alpha(n+1) \rfloor$ and the upper endpoint the k^+-th smallest with $k^+ = \lceil (1-\alpha)(n+1) \rceil$. This yields a provable coverage guarantee of at least $1 - 2\alpha$ irrespective of the data distribution or algorithmic stability, at the cost of retaining all leave-one-out predictions at the test points [14], [15]. Both guarantees rest on exchangeability of the training and test observations, an assumption examined in Section 3.9 [37].

2.6 Simulation Design

Monte Carlo simulation evaluates a statistical procedure under conditions the analyst controls [38]. To be relevant to applied work, the generated data must resemble the empirical data in both the shape of each marginal distribution and the dependence among variables [39]. A semiparametric approach met both requirements: dependence was modeled parametrically by a Gaussian copula, each marginal nonparametrically by its empirical distribution.

By Sklar's theorem [40], any d-dimensional joint distribution function can be decomposed as

$$F(x_1, \dots, x_d) = C\big(F_1(x_1), \dots, F_d(x_d)\big) \tag{10}$$

with F_j the marginals and C a copula, that is, a joint distribution with uniform marginals [41], [42]. The Gaussian copula, obtained from the multivariate normal distribution with correlation matrix R, is

$$C_R^{\text{Gauss}}(u) = \Phi_R\big(\Phi^{-1}(u_1), \dots, \Phi^{-1}(u_d)\big) \tag{11}$$

Dependence was estimated on the normal-score scale. Each observation was transformed to a van der Waerden score,

$$z_{ij} = \Phi^{-1}\left(\frac{\text{rank}(x_{ij})}{n+1}\right) \tag{12}$$

with the analogous construction for categorical variables via cumulative category proportions and interval midpoints. The correlation matrix R was estimated from these scores and projected onto the nearest positive-definite correlation matrix where necessary [43], followed by a Cholesky decomposition $R = LL^{\top}$. Synthetic predictors were then generated as

$$Z = \xi L^{\top}, \quad U = \Phi(Z), \quad x_j^* = \hat{F}_j^{-1}\big(U_j\big) \tag{13}$$

with ξ a standard normal vector, so that $Z \sim N(0, R)$. For numeric variables, $\hat{F}_j^{-1}$ was the empirical quantile function, applied through inverse transform sampling [44], [36]; for categorical variables, U_j was mapped to a category through the empirical cumulative proportions. Constant variables were fixed at their observed value. The procedure therefore preserved both the empirical marginal of each variable, including the pronounced skewness of the numeric ones, and the rank dependence among them.

The choice of a Gaussian copula requires justification, since it preserves rank dependence while imposing a specific dependence structure. It extends naturally to the dimension required here (82 columns), is estimable from a moderate sample, and, combined with empirical marginals, leaves the univariate shapes untouched. Its principal limitation is zero tail dependence, so extreme values of two variables become asymptotically independent. This matters for expenditure data, where jointly extreme outcomes are precisely the configurations in which the intervals were later found to fail. Section 3.2 quantifies how closely the synthetic data reproduced the marginals and the correlation structure of the source data, benchmarked against a random subsample of the same size.

The response was generated from a random forest fitted to the empirical data, with additive Gaussian noise:

$$Y_i = \widehat{f_{\text{RF}}}(X_i) + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad \sigma = 0.10\, s_Y \tag{14}$$

Because the response-generating function and noise level were fixed by construction, the X-Y relationship was realistic yet fully controlled, permitting systematic assessment of coverage and width across dimensionality regimes.

Ten scenarios were defined by a target ratio $r = p/n$, where p = 67 denotes the number of original explanatory variables (82 columns after encoding) and n the training sample size, $n = \text{round}\left(\frac{p}{r}\right)$ with test size $m = \text{round}(0.25n)$ corresponding to an 80:20 split of the generated dataset. Each scenario was replicated 100 times, giving 1,000 datasets per method-model combination and 6,000 in total. Table 4 lists the scenarios.

Table 4. Simulation scenarios by $\frac{p}{n}$ ratio (p = 67 original variables)

No.	$\frac{p}{n}$	Training (n)	Test (m)	No.	$\frac{p}{n}$	Training (n)	Test (m)
1	0.2	335	84	6	1.2	56	14
2	0.4	168	42	7	1.4	48	12
3	0.6	112	28	8	1.6	42	10
4	0.8	84	21	9	1.8	37	9
5	1.0	67	17	10	2.0	34	8

2.7 Evaluation Metrics

Interval quality was assessed on five metrics, with m the number of test observations. The prediction coverage rate (PCR) measures reliability:

$$\text{PCR} = \frac{1}{m}\sum_{j=1}^{m} 1\{L_j \le Y_{n+j} \le U_j\} \tag{15}$$

An interval is valid when empirical coverage reaches the nominal level. The average interval width (AIW) measures sharpness:

$$\text{AIW} = \frac{1}{m}\sum_{j=1}^{m}\left(U_j - L_j\right) \tag{16}$$

Width cannot be interpreted on its own, since a very narrow interval will fail to cover; PCR and AIW represent a trade-off and must be read together [45]. The Winkler score combines both into a single proper scoring rule [46], [47]:

$$WS_j = \left(U_j - L_j\right) + \frac{2}{\alpha}\left(L_j - Y_{n+j}\right)1\{Y_{n+j} < L_j\} + \frac{2}{\alpha}\left(Y_{n+j} - U_j\right)1\{Y_{n+j} > U_j\} \tag{17}$$

with the reported score being the mean over test observations. Computation time (CT, in seconds) and peak memory usage (MU, in MB) complete the set. Lower values are preferred for AIW, WS, CT, and MU, whereas PCR should be as close as possible to the nominal 90% from above. All width and score figures are reported in millions of rupiah (M IDR).

2.8 Software, Measurement Protocol, and Code Availability

All analyses were conducted in Python 3.10 on Google Colab using NumPy, pandas, SciPy, scikit-learn, and Matplotlib, on a 2-core Intel Xeon CPU (2.20 GHz) with 12.7 GB RAM and approximately 100 GB of temporary storage. We measured computation time using time.perf_counter over the full interval construction, including all leave-one-out refits, the final fit where applicable, and test-set prediction, excluding data loading and preprocessing. Estimators ran single-threaded for comparability. We measured peak Python-level memory allocation using tracemalloc, rather than process resident memory (Section 3.7). Absolute timings may vary across hardware. We deterministically assigned random seeds by replication index to ensure reproducibility across all 6,000 runs. Analysis scripts, seeds, package versions, and reproduction instructions are available at https://github.com/meetaaa/jackknife-. The repository contains code only; Susenas microdata are not redistributed due to Statistics Indonesia access restrictions. The README specifies the required source files, variable definitions, and directory structure, enabling replication once you obtain the microdata through the official request procedure.

3. RESULT AND ANALYSIS

3.1 Data Description and Preprocessing Outcomes

Household monthly expenditure was severely right-skewed (Table 5): the mean greatly exceeded the median, the range spanned 0.73 to 316.83 M IDR, and the skewness of 11.16 reflected a long right tail generated by a small number of very high-spending households. Under the 1.5 x interquartile range rule, 93 observations (11.4%) were flagged as outliers. These were retained, since they reflect genuine heterogeneity in welfare rather than measurement error, a pattern typical of expenditure and income data [26]. A natural-log transformation reduced skewness from 11.16 to 0.73 and was applied for support vector regression.

Table 5. Descriptive statistics of the response variable EXPEND (M IDR per month)

N	Mean	Median	SD	Minimum	Maximum	Skewness	IQR outliers
817	9.69	5.08	15.30	0.73	316.83	11.16	93 (11.4%)

Among the numeric predictors, floor area (R1804) was continuous and right-skewed (skewness 4.01), ranging from 3 to 900 m² with a mean of 78.58 m², and showed the strongest association with expenditure (Spearman $\rho = 0.527$). The four household-size counts (R302-R305) were small and discrete and correlated only weakly with expenditure (ρ from -0.064 to 0.147). R303 and R304 were strongly collinear ($\rho = 0.894$), which is conceptually expected because members aged 10 and over are a subset of members aged 5 and over; this does not impair prediction for the models used but is noted because the two variables may offset one another in any importance analysis.

Asset ownership was the most informative block of binary indicators, consistent with the use of asset indices as welfare proxies [20]: refrigerators were owned by 81.3% of households, motorcycles by 72.5%, and land by 65.7%, whereas higher-value assets such as air conditioners (18.4%) and cars (18.0%) were much rarer, with an average of 4.19 of the 13 recorded asset types per household. In the food insecurity block, which adopts the Food Insecurity Experience Scale, 184 households (22.5%) reported at least one symptom. Log-expenditure separated sharply by high-value asset ownership: median expenditure was 19.40 M IDR for air conditioner

owners against 4.52 M IDR for non-owners, and 20.48 M IDR against 4.52 M IDR for car owners. After preprocessing, the analysis dataset comprised 815 households and 82 encoded columns.

3.2 Simulation: Fidelity of the Synthetic Data

Before interpreting the interval results, the generator itself requires verification. Five discrepancy measures were computed in each replication against the full empirical dataset: the Kolmogorov-Smirnov statistic KS_{EXP} and the normalized Wasserstein distance (W_{EXP}) for the response, the mean Kolmogorov-Smirnov statistic across the numeric predictors KS_{num}, the mean total variation distance across the categorical predictors (TV_{cat}), and the mean absolute difference in Spearman correlations Δ_{corr}. Because any sample of size n differs from the full data by chance alone, each measure was reported alongside a baseline obtained by drawing a random subsample of the same size from the empirical data. That baseline is the discrepancy attributable to sample size rather than to the generator, and it is the appropriate comparison. Table 6 reports both, averaged over 100 replications.

Table 6. Mean discrepancy of the generated data (b) and of an empirical-subsample baseline (d) relative to the full empirical data, by $\frac{p}{n}$ ratio (mean of 100 replications)

$\frac{p}{n}$	n	$KS_{EXP,b}$	$KS_{EXP,d}$	$W_{\text{EXP},b}$	$W_{EXP,d}$	$KS_{num,b}$	$KS_{num,d}$	$TV_{cat,b}$	$TV_{cat,d}$	$\Delta_{\text{corr},b}$	$\Delta_{\text{corr},d}$
0.2	335	0.1952	0.0363	0.1535	0.0584	0.0286	0.0217	0.0106	0.0081	0.0647	0.0299
0.4	168	0.2067	0.0582	0.1605	0.0889	0.0409	0.0368	0.0149	0.0135	0.0755	0.0494
0.6	112	0.2080	0.0746	0.1649	0.1034	0.0517	0.0473	0.0185	0.0171	0.0846	0.0625
0.8	84	0.2104	0.0876	0.1710	0.1252	0.0585	0.0543	0.0211	0.0197	0.0933	0.0763
1.0	67	0.2181	0.0964	0.1797	0.1325	0.0655	0.0603	0.0229	0.0221	0.1052	0.0855
1.2	56	0.2218	0.1096	0.1843	0.1382	0.0701	0.0676	0.0246	0.0251	0.1071	0.0919
1.4	48	0.2177	0.1208	0.1858	0.1691	0.0754	0.0756	0.0273	0.0280	0.1179	0.1008
1.6	42	0.2279	0.1259	0.1907	0.1746	0.0824	0.0794	0.0291	0.0291	0.1250	0.1052
1.8	37	0.2313	0.1346	0.1947	0.1619	0.0880	0.0832	0.0313	0.0311	0.1320	0.1146
2.0	34	0.2392	0.1419	0.1964	0.1782	0.0909	0.0845	0.0325	0.0323	0.1392	0.1180

Note: KS_{EXP} = Kolmogorov-Smirnov statistic for EXPEND; W_{EXP} = normalised Wasserstein distance for EXPEND; KS_{num}= mean Kolmogorov-Smirnov statistic over the numeric predictors; TV_{cat} = mean total variation distance over the categorical predictors; Δ_{corr} = mean absolute difference in Spearman correlations. Lower values indicate closer agreement with the full empirical data.

For the explanatory variables, the generator performed essentially as well as random subsampling of the same size. The mean Kolmogorov-Smirnov statistic for the numeric predictors never departed from its baseline by more than about 0.007, and from p/n = 1.4 onward the two were indistinguishable (0.0754 against 0.0756 at that ratio). The categorical total variation distance behaved the same way, with a gap that never exceeded 0.0025 and that reversed sign at the higher ratios. Rank dependence was preserved less exactly: Δ_{corr} exceeded its baseline at every ratio, by 0.035 at p/n = 0.2, although the gap narrowed to 0.021 at p/n = 2.0 as sampling variation came to dominate.

The response was the one component that departed substantially from its baseline (KS_{EXP} of 0.1952 against 0.0363 at p/n = 0.2). This is a consequence of the design rather than a defect of the copula: the synthetic response was not resampled but generated from the fitted random forest with additive noise (Eq. 14). Hence, its distribution is by construction a smoothed version of the empirical one, with the extreme right tail attenuated. Both KS_{EXP} and W_{EXP} were stable across ratios, confirming that the attenuation is a fixed property of the response model rather than something that worsens as n shrinks. Because the empirical coverage failures documented in Section 3.8 concentrate precisely in that attenuated tail, the simulation understates how often such failures arise.

3.3 Simulation: Comparison of Jackknife and Jackknife+

Across all scenarios and replications, every method-model combination produced a median coverage slightly above the nominal 90%, around 91-92%, and the jackknife+ consistently attained a marginally higher median than the jackknife for all three base models.

The aggregate medians, however, conceal substantial variability. The first quartile of the coverage distribution lay below the nominal level for every combination, at roughly 86-88%, so nominal coverage was not achieved in at least a quarter of the replications. The distributions were left-skewed: upper whiskers typically reached 100% while lower whiskers descended to about 64-80%, and outliers below the lower whisker, between 47% and 70%, occurred for every model and method. In those replications the interval covered only about half to two-thirds of the actual values, a serious shortfall relative to the 90% target. The jackknife+ showed a slightly

higher first quartile and a shorter lower tail, indicating a marginally lower risk of under-coverage, the direction the theory predicts.

Interval width behaved almost identically under the two methods: across all three base models the width distributions overlapped closely in median, quartiles, and whisker length, so the choice between jackknife and jackknife+ did not materially change the sharpness of the intervals. Winkler scores were correspondingly similar, the jackknife+ attaining a slightly lower median for all three models. Since the Winkler score penalizes non-coverage, this small and consistent advantage arises from a reduced penalty component rather than from narrower intervals.

The two methods differed most clearly in computational cost. The jackknife+ was equal to or slightly slower than the jackknife in every configuration, the additional burden arising from evaluating each leave-one-out model at every test point and assembling the candidate endpoints. Memory usage showed the same direction with a larger gap in the upper tail: median usage was only slightly higher for the jackknife+, but its upper quartile, upper whisker, and maximum outliers were consistently above those of the jackknife.

3.4 Simulation: Effect of the $\frac{p}{n}$ Ratio

Increasing p/n, which in this design means shrinking the training sample while holding the number of predictors fixed, produced two opposing consequences: statistically, intervals widened, Winkler scores rose, and coverage became more volatile; computationally, time and memory fell because fewer observations were processed.

Average interval width increased steadily with p/n, following from the reduced information available for estimating the predictor-response relationship: as p approached and exceeded n, the model became more sensitive to sample variation, and the additional predictive uncertainty had to be absorbed by wider intervals. Winkler scores rose correspondingly, with a pronounced spike at p/n = 1.6 where intervals both widened and incurred non-coverage penalties, followed by a decline at 1.8 and a further rise at 2.0.

Coverage, by contrast, showed no monotone trend. Most values lay at or slightly above 90%, around 90-91% at ratios 0.2 and 0.4 and roughly 91-94% at ratios 1.2-1.6, with irregular dips at higher ratios. Sample size alone therefore does not determine coverage; what changes with p/n is the volatility of coverage rather than its central tendency. Computation time and memory declined monotonically as p/n rose, and the gap between the two interval methods narrowed accordingly, since the total workload itself shrank.

Three regimes emerged. At low ratios (0.2-0.6), intervals were narrow, Winkler scores low to moderate, coverage not always stable, and computational cost very high. At intermediate ratios (0.8-1.2), intervals began to widen, coverage was generally good, Winkler scores rose, and computational cost fell sharply. At high ratios (1.4-2.0), statistical uncertainty was large, width and Winkler scores were high, coverage became more sensitive to the particular combination of model and method, and computational cost was low.

3.5 Simulation: Effect of the Base Model

The three base models displayed distinct profiles, summarised in Table 7. Random forest delivered the best statistical interval quality: coverage generally at or above 90% apart from a dip at p/n = 0.6, width rising only gradually from about 9-11 M IDR at low ratios to roughly 18 M IDR at p/n = 2.0, and the lowest Winkler score at most ratios. Its weakness was speed: the slowest of the three, requiring approximately 244-258 s at p/n = 0.2 and about 22-23 s at p/n = 2.0.

Gradient boosting was close to random forest in interval quality while being far cheaper. Its widths matched random forest at ratios 0.2-1.0 (roughly 9-14 M IDR) but rose more steeply at the extreme, reaching about 23 M IDR at p/n = 2.0, near the level of support vector regression. Its computation time fell from about 61 s at p/n = 0.2 to about 5 s at p/n = 2.0, and it was the most memory-efficient of the three, usage becoming very low and stable once p/n exceeded roughly 0.8.

Support vector regression behaved differently from the two tree ensembles. Its coverage was adequate but was obtained through markedly wider intervals: about 12.7 M IDR at p/n = 0.2, roughly 19 M IDR at p/n = 1.0, and above 22 M IDR at high ratios, the largest at nearly every ratio. Its Winkler score was correspondingly the highest, peaking at about 45.5 M IDR for the jackknife at p/n = 1.6, so the extra width did not buy proportionate reliability. Its advantage was speed: the fastest at every ratio, from about 17-20 s at p/n = 0.2 down to about one second at p/n = 2.0. That advantage did not extend to memory, which was comparatively high at low ratios, particularly with the jackknife+.

Table 7. Profile of the three base models in the simulation study

Model	Interval width	Winkler score	Computation time	Memory
Random forest	Lowest; rises gradually (9-11 to 18 M IDR)	Lowest at most ratios	Highest (244-258s to 22-23s)	Intermediate
Gradient boosting	Comparable to RF; steeper rise at $p/n = 2.0$ (23 M IDR)	Competitive with RF	Moderate (61s to 5s)	Lowest
Support vector regression	Highest at nearly all ratios (12.7 to >22 M IDR)	Highest (peak 45.5 M IDR)	Lowest (17-20s to 1s)	High at low ratios

3.6 Empirical Application: Point Prediction Performance

The preprocessed data were split 80:20 into 652 training and 163 test households. Table 8 reports root mean squared error and the coefficient of determination on the held-out test set alongside five-fold cross-validated values, the latter included because a single random split can be unstable at this sample size.

Table 8. Model performance on the test set and under five-fold cross-validation

Model	RMSE (M IDR)	RMSE-CV (M IDR)	R^2	R^2-CV
Gradient boosting	6.99	12.59	0.564	0.401
Random forest	6.72	12.37	0.597	0.422
Support vector regression	6.32	13.70	0.643	0.291

Note: CV denotes five-fold cross-validation. RMSE values are in millions of rupiah (M IDR).

The two evaluations disagreed, and the disagreement is itself informative. On the single test split, support vector regression performed best (RMSE 6.32 M IDR, R^2 0.643), followed by random forest and gradient boosting. Under cross-validation the ranking reversed: random forest was best (RMSE 12.37 M IDR, R^2 0.422), gradient boosting second, and support vector regression last (RMSE 13.70 M IDR, R^2 0.291). The apparent superiority of support vector regression on one partition therefore did not survive a change in the training and validation sets, whereas random forest generalized most consistently across folds. This is the clearest evidence in the study that conclusions from a single 80:20 split at this sample size should be treated with caution.

Cross-validated RMSE exceeded test-set RMSE for all three models, attributable to the extreme skewness of the response: when households with extreme expenditure fall unevenly across folds, the error on the affected fold rises sharply. Scatter plots of actual against predicted values showed points clustering along the identity line in the low and middle expenditure range but diverging progressively in the upper range, the largest prediction errors, roughly 48 to 59 M IDR, occurring for households approaching 80 M IDR in expenditure, all systematically underpredicted. Residual plots displayed a funnel shape, residual variance growing as the fitted value increased, indicating heteroscedasticity. This matters for what follows, because the jackknife construction assumes a single global residual quantile.

3.7 Empirical Application: Prediction Interval Performance

Table 9 reports the five metrics for all six method-model combinations on the empirical data at the nominal 90% level.

Table 9. Prediction interval performance on the empirical data (n = 652 training, m = 163 test; nominal coverage 90%)

Model	Method	PCR (%)	AIW (M IDR)	WS (M IDR)	CT (s)	MU (MB)
Gradient boosting	Jackknife	92.02	18.19	29.85	156.28	0.75
Gradient boosting	Jackknife+	92.02	18.27	29.82	155.28	5.13
Random forest	Jackknife	91.41	17.76	30.89	638.44	1.10
Random forest	Jackknife+	90.80	18.02	30.25	660.31	5.34
Support vector regression	Jackknife	92.64	20.43	31.04	88.81	1.99
Support vector regression	Jackknife+	92.64	20.43	30.93	103.08	5.26

All six combinations exceeded the nominal level, with coverage between 90.80% and 92.64%; none under-covered. Random forest with the jackknife+ was closest to nominal (90.80%, a deviation of +0.80 percentage points), followed by random forest with the jackknife (91.41%, +1.41). The two interval methods differed only for random forest: gradient boosting yielded 92.02% under both and support vector regression 92.64% under both. This is consistent with the theoretical distinction between the methods, since the jackknife+ differs from the jackknife precisely when the fitted function is sensitive to the removal of a single observation, and random forest is the model displaying that sensitivity here.

Because the Winkler score is the sum of interval width and a non-coverage penalty, the difference between WS and AIW isolates the penalty each configuration incurs (Table 10). The decomposition inverts the impression given by width alone: random forest produced the narrowest intervals but the largest penalty (12.23-13.13 M IDR), while support vector regression produced the widest intervals but the lightest penalty (10.50-10.61 M IDR). Gradient boosting was intermediate on both components and consequently obtained the best combined score. The jackknife+ attained a lower penalty than the jackknife for all three base models, the only respect in which its advantage was consistent across models.

Table 10. Decomposition of the Winkler score into interval width and non-coverage penalty (M IDR)

Model	Method	AIW	WS	Penalty
Gradient boosting	Jackknife+	18.27	29.82	11.55
Gradient boosting	Jackknife	18.19	29.85	11.66
Random forest	Jackknife+	18.02	30.25	12.23
Random forest	Jackknife	17.76	30.89	13.13
Support vector regression	Jackknife+	20.43	30.93	10.50
Support vector regression	Jackknife	20.43	31.04	10.61

Whereas the interval quality metrics differed by less than 4% across configurations, computation time varied by more than sixfold. Random forest was the most expensive at 660.31 s, the jackknife+ costing about 4.3 times as much as gradient boosting (155.28 s) and 6.4 times as much as support vector regression (103.08 s). The additional time required by the jackknife+ over the jackknife was small for random forest (+3.4%), effectively nil for gradient boosting, and largest in relative terms for support vector regression (+16.1%). Memory showed the largest relative differences: the jackknife+ required 2.6 to 6.8 times the memory of the jackknife for the same model. More informative is the uniformity of jackknife+ consumption across models (5.13-5.34 MB) against the variability of jackknife consumption (0.75-1.99 MB). This follows from the structure of the two procedures: jackknife+ memory is dominated by the matrix of leave-one-out predictions at the test points, sized by the sample and not by model complexity, whereas the jackknife stores only n residuals plus the fitted model, so its footprint tracks the individual algorithm. The memory advantage of the jackknife therefore grows with sample size, the relevant consideration for deployment on datasets larger than the one analyzed here.

3.8 Diagnostics: Where the Intervals Fail

Plotting the 163 test intervals observation by observation revealed three features that the aggregate metrics conceal, and these are more consequential for practice than the small differences between methods reported above. First, interval width was nearly uniform across observations: a household with a low predicted expenditure received an interval of almost the same width as one with a high predicted expenditure. This follows directly from the jackknife construction, which adds and subtracts the same quantile of absolute leave-one-out residuals to every prediction, so width does not depend on the predicted value [15]. Where the errors are heteroscedastic, as Section 3.6 established for these data, this property is poorly matched to the problem: intervals were unnecessarily wide for the majority of lower-spending households and too narrow for the higher-spending ones. Second, the lower bound was negative for a large proportion of observations. Since household expenditure cannot be negative, that portion of the interval carries no information and inflates the reported width without contributing to coverage; truncating at zero, or modeling on the log scale throughout rather than for support vector regression alone, would address this, and both are natural refinements for subsequent work. Third, coverage failures were concentrated in the right tail. Observations well above the median were systematically excluded from the intervals; a household with expenditure near 84 M IDR, for example, was covered by none of the six combinations. Because the Winkler score penalizes in proportion to the distance from the nearest endpoint, this small group of tail observations contributed the bulk of the penalty component in Table 10. The comparison across base models explains the counter-intuitive penalty ordering: support vector regression shrinks large observations least, so some tail cases fell inside its intervals, whereas gradient boosting and random forest predicted much lower values for those same households and failed to cover them. The advantage of support vector regression on the penalty component therefore reflects a wider spread of predictions rather than greater accuracy, and that same spread produces its larger overall width.

The visual comparison of the two interval methods reinforced the aggregate findings. For gradient boosting and support vector regression the two sets of intervals were practically indistinguishable, consistent with their identical PCR and AIW values; for random forest the endpoints differed visibly for several observations, confirming the sensitivity of that model to the removal of a single training point.

3.9 Limitations

Several limitations qualify the findings. The unweighted analysis of Susenas data does not account for its stratified multistage design, limiting coverage interpretation to the 815 surveyed households rather than the Bogor City population [7], [15]. Generalizability is further limited by the use of one region, one survey year, and a single 80:20 split, despite repeated simulations and cross-validation. Small coverage differences should therefore be interpreted cautiously. Additionally, the Gaussian copula does not capture joint tail dependence, fixed dimensionality introduces increasing sparsity as sample size decreases, and hyperparameters and computational conditions were not varied. Consequently, the findings apply primarily to the relative performance of jackknife and jackknife+ under the evaluated conditions.

4. CONCLUSION

This study compared jackknife and jackknife+ conformal prediction intervals for random forest, gradient boosting, and support vector regression, in a simulation across ten p/n ratios and in an application to Susenas 2023 household expenditure data for Bogor City. In the simulation both methods produced median coverage near the nominal level, but the first quartile fell below it and the distributions were left-skewed, a minority of replications covering as little as 47-70% of test observations. The jackknife+ consistently attained a slightly higher median coverage, a shorter lower tail, and a marginally lower Winkler score, at the cost of modestly higher computation time and appreciably higher memory use. Raising the p/n ratio widened the intervals and increased Winkler scores while lowering computational cost; coverage showed no monotone trend but became more volatile. In the empirical application to 815 households, all six combinations covered between 90.80% and 92.64% of test observations, so none under-covered. Random forest with the jackknife+ came closest to nominal, while gradient boosting with the jackknife+ achieved the lowest Winkler score at roughly one quarter of the computation time of random forest. No single combination was best on all five metrics.

Two qualifications are essential to the correct reading of these results. First, the observation-level diagnostics in Section 3.8 show that the apparent over-coverage is not uniform: high average coverage here is compatible with poor interval quality for exactly the households whose expenditure is hardest to predict. Second, the closeness of the two methods in interval quality, and the modest advantage of the jackknife+ in the Winkler score, are specific to this dataset, this region, this survey year, and these model configurations. They are not universal claims about the relative merits of jackknife and jackknife+, and they say nothing about adaptive conformal methods, which were not evaluated here.

Subject to those qualifications, the jackknife+ is the preferable choice for expenditure modeling of the kind studied here, since it improves the Winkler score for all three base models and yields the coverage closest to nominal for random forest, at a small additional cost in time. Its memory requirement, however, grows with the sample size by construction, so deployment on substantially larger datasets requires deliberate resource planning, and the jackknife remains the sensible option where memory is the binding constraint. Extending the analysis to further regions, additional nominal levels, and a wider range of p/n ratios, and comparing both methods against interval constructions that adapt to heteroscedasticity, are the priorities identified for subsequent work.

ACKNOWLEDGEMENT

The authors thank Statistics Indonesia for providing access to the Susenas 2023 microdata. This research received no specific grant from any funding agency

5. REFERENCES

[1] L. Breiman, "Random Forest," *Mach. Learn.*, vol. 45, pp. 1–33, 2001.

[2] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, pp. 1189–1232, 2001, https://doi.org/10.1214/aos/1013203451.

[3] V. Vapnik, *The Nature of Statistical Learning Theory, Second Edition*. 1995.

[4] H. Drucker, C. J. C. Surges, L. Kaufman, A. Smola, and V. Vapnik, "Support vector regression machines," *Adv. Neural Inf. Process. Syst.*, vol. 1, pp. 155–161, 1997.

[5] T. Hastie, R. Tibshirani, and J. Friedman, "Springer Series in Statistics The Elements of Statistical Learning," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2009.

[6] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning with Applications in R*, vol. 7, no. 10. New York: Springer, 2013. https://doi.org/10.1007/978-1-4614-7138-7.

[7] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. New York: Springer, 2005.

[8] G. Shafer and V. Vovk, "A tutorial on conformal prediction," *Journal of Machine Learning Research*, vol. 9, pp. 371–421, 2008.

[9] H. Papadopoulos, K. Proedrou, V. Vovk, and A. Gammerman, "Inductive Confidence Machines for Regression BT - Machine Learning: ECML 2002," T. Elomaa, H. Mannila, and H. Toivonen, Eds., Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 345–356.

[10] J. Lei, M. G'Sell, A. Rinaldo, R. J. Tibshirani, and L. Wasserman, "Distribution-Free Predictive Inference for Regression," *J. Am. Stat. Assoc.*, vol. 113, no. 523, pp. 1094–1111, 2018, https://doi.org/10.1080/01621459.2017.1307116.

[11] Y. Romano, E. Patterson, and E. J. Candès, "Conformalized quantile regression," *Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[12] C. Kath and F. Ziel, "Conformal prediction interval estimation and applications to day-ahead and intraday power markets," *Int. J. Forecast.*, vol. 37, no. 2, pp. 777–799, 2021, https://doi.org/10.1016/j.ijforecast.2020.09.006.

[13] A. N. Angelopoulos and S. Bates, "A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification," 2022.

[14] M. Fontana, G. Zeni, and S. Vantini, "Conformal prediction: A unified review of theory and new challenges," *Bernoulli*, vol. 29, no. 1, pp. 1–23, 2023, https://doi.org/10.3150/21-BEJ1447.

[15] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Predictive inference with the jackknife+," *Ann. Stat.*, vol. 49, no. 1, pp. 486–507, 2021, https://doi.org/10.1214/20-AOS1965.

[16] L. Steinberger and H. Leeb, "Conditional predictive inference for high-dimensional stable algorithms," pp. 1–75, 2018.

[17] A. N. Angelopoulos, R. F. Barber, and S. Bates, "Theoretical Foundations of Conformal Prediction," 2026.

[18] V. Manokhin, "Machine Learning for Probabilistic Prediction," Royal Holloway University of London, 2022.

[19] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon, "Combining satellite imagery and machine learning to predict poverty," *Science (1979).*, vol. 353, no. 6301, pp. 790–794, 2016, https://doi.org/10.1126/science.aaf7894.

[20] D. Filmer and L. H. Pritchett, "Estimating wealth effects without expenditure data—or tears: An application to educational enrollments in states of India," *Demography*, vol. 38, no. 1, pp. 115–132, Feb. 2001, https://doi.org/10.1353/dem.2001.0003.

[21] E. L. Aiken, G. Bedoya, J. E. Blumenstock, and A. Coville, "Program targeting with machine learning and mobile phone data: Evidence from an anti-poverty intervention in Afghanistan," *J. Dev. Econ.*, vol. 161, p. 103016, 2023, https://doi.org/10.1016/j.jdeveco.2022.103016.

[22] A. Usmanova, A. Aziz, D. Rakhmonov, and W. Osamy, "Utilities of Artificial Intelligence in Poverty Prediction: A Review," *Sustainability (Switzerland)*, vol. 14, no. 21, pp. 1–39, 2022, https://doi.org/10.3390/su142114238.

[23] BPS Kota Bogor, "Kota Bogor dalam Angka 2023," Kota Bogor, Indonesia, 2023.

[24] Badan Pusat Statistik, *Pengeluaran untuk Konsumsi Penduduk Indonesia*. Jakarta, 2023. https://doi.org/10.21831/jptk.v23i1.10441.
[25] Badan Pusat Statistik, *Statistik Kesejahteraan Rakyat 2023*. Jakarta: [BPS] Badan Pusat Statistik, 2023.
[26] A. Deaton and S. Zaidi, "Guidelines for constructing consumption aggregates for welfare analysis," in *World Bank LSMS Working Paper 135*, Washington, DC, USA: The International Bank for Reconstruction and Development / The World Bank, 2002.
[27] S. L. Lohr, *Design and Analysis*. Duxburry Press, 2022.
[28] R. J. A. Little and D. B. Rubin, *Statistical Analysis with Missing Data*, Third edit. New Delhi, India: Wiley, 2020.
[29] T. de Waal, J. Pannekoek, and S. Scholtus, "Handbook of Statistical Data Editing and Imputation," 2011.
[30] M. Kuhn and K. Johnson, *Applied Predictive Modeling*. Groton, Connecticut, USA: Springer, 2013.
[31] C. Cortes and V. Vapnik, "Support Vector Network," *Mach. Learn.*, vol. 20, pp. 273–297, 1995, https://doi.org/10.1007/BF00994018.
[32] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Stat. Comput.*, vol. 14, no. 3, pp. 199–222, 2004, https://doi.org/10.1023/B:STCO.0000035301.49549.88.
[33] M. H. Quenouille, "Approximate Tests of Correlation in Time-Series," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 11, no. 1, pp. 68–84, Jan. 1949, https://doi.org/10.1111/j.2517-6161.1949.tb00023.x.
[34] John W. Tukey, "Bias and Confidence in Not-quite Large Samples," *The Annals of Mathematical Statistics*, vol. 29, p. 614, 1958.
[35] R. G. Miller, "The Jackknife--A review," 1974.
[36] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics*, vol. 7, no. 1, pp. 1–26, 1979.
[37] R. F. Barber, E. J. Candès, A. Ramdas, and R. J. Tibshirani, "Conformal prediction beyond exchangeability," *The Annals of Statistics*, vol. 51, no. 2, pp. 816–845, 2023, https://doi.org/10.1214/23-AOS2276.
[38] N. Metropolis and S. Ulam, "The Monte Carlo Method," *J. Am. Stat. Assoc.*, vol. 44, no. 247, pp. 335–341, Sep. 1949, https://doi.org/10.1080/01621459.1949.10483310.
[39] B. Nowok, G. M. Raab, and C. Dibben, "synthpop: Bespoke Creation of Synthetic Data in R," *J. Stat. Softw.*, vol. 74, no. 11 SE-Articles, pp. 1–26, Oct. 2016, https://doi.org/10.18637/jss.v074.i11.
[40] M. Sklar, "Fonctions de répartition à N dimensions et leurs marges To cite this version : HAL Id : hal-04094463," *hal.scienceM SklarAnnales de l'ISUP, 1959 • hal.science*, no. 3, pp. 229–231, 1959.
[41] R. B. Nelsen, *An Introduction to Copulas*. New York: Springer, 2006. https://doi.org/10.1007/0-387-28678-0.
[42] H. Joe, "Dependence modeling with copulas," *Dependence Modeling with Copulas*, pp. 1–457, 2014, https://doi.org/10.1201/b17116.
[43] N. J. Higham, "Computing the nearest correlation matrix—a problem from finance," *IMA Journal of Numerical Analysis*, vol. 22, no. 3, pp. 329–343, Jul. 2002, https://doi.org/10.1093/imanum/22.3.329.
[44] L. Devroye, "Non-Uniform Random Variate Generation," *Springer-Verlag, New York*, 1986, https://doi.org/10.1007/978-1-4613-8643-8.
[45] A. Khosravi, S. Nahavandi, D. Creighton, and A. F. Atiya, "Comprehensive review of neural network-based prediction intervals and new advances.," *IEEE Trans. Neural Netw.*, vol. 22, no. 9, pp. 1341–1356, Sep. 2011, https://doi.org/10.1109/TNN.2011.2162110.
[46] R. L. Winkler, "A decision-theoretic approach to interval estimation," *J. Am. Stat. Assoc.*, vol. 67, no. 337, pp. 187–191, 1972, https://doi.org/10.1080/01621459.1972.10481224.
[47] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007, https://doi.org/10.1198/016214506000001437.