Simulation-Augmented Multi-Step Split Conformal Prediction for Aggregated Forecasts

Anonymous Authors[1]

Abstract

We study uncertainty quantification for aggregated forecasting tasks such as annual totals and year-over-year growth rates. We propose SA-MSCP, a simulation-augmented multi-step split conformal method that generates future paths from cross-validated residuals using a block bootstrap and constructs prediction intervals from empirical quantiles. Experiments show that SA-MSCP improves empirical coverage over a simulated-path baseline for aggregated and growth-rate targets. Our results demonstrate that simulation-enhanced conformal calibration is an effective and general framework for uncertainty quantification in aggregated time-series forecasting.

1. Introduction

Prediction intervals for time-series forecasts are central to applications such as retail and finance. Conformal prediction (CP) provides distribution-free coverage under exchangeability (Vovk et al., 2005; Lei et al., 2018). Extensions address temporal dependence (Xu & Xie, 2021; 2023; Wang & Hyndman, 2024), online adaptation (Gibbs & Candès, 2021; Angelopoulos et al., 2023), weighted quantiles (Tibshirani et al., 2019; Barber et al., 2023), time-series datasets (Stankevičiūtė et al., 2021; Sun & Yu, 2022), and distribution shifts (Zou & Liu, 2024).

Many applications require uncertainty for aggregated targets such as annual sums and year-over-year growth rates. This is challenging because aggregation induces dependence and nonlinear transformations, so combining pointwise intervals is invalid, and CP guarantees do not directly extend to multi-step dependent settings.

We propose SA-MSCP, a simulation-enhanced multi-step split conformal method (Wang & Hyndman, 2024) for aggregated forecasting. The method uses expanding-window cross-validation to collect residuals across forecast origins, then simulates future paths with a block bootstrap over residual sequences to better preserve dependence before constructing prediction intervals from empirical quantiles of aggregated trajectories.

Related work includes copula-based calibration for joint coverage (Sun & Yu, 2022), conformalized interval arithmetic for sums under group exchangeability (Luo & Zhou, 2025), and hierarchical reconciliation approaches (Principato et al., 2025). In contrast, SA-MSCP focuses on temporal aggregation within a univariate series.

Our contributions are: (i) a practical MSCP-based procedure for aggregated targets that improves empirical coverage without finite-sample guarantees, and (ii) an empirical evaluation on M4 and a proprietary dataset of 2,000 series showing consistent coverage gains over a simulated-path baseline with significance assessed via Wilcoxon tests.

2. Methodology

Split conformal prediction (SCP) constructs prediction intervals from calibration residuals and achieves marginal coverage under exchangeability (Lei et al., 2018; Shafer & Vovk, 2008). In time series, temporal dependence violates this assumption, motivating multi-step conformal approaches that approximate residual distributions via rolling or expanding windows.

SA-MSCP follows this framework. Residuals $\hat{\varepsilon}_{k,h} = y_{k+h} - \hat{y}_{k+h}$ are collected via expanding-window cross-validation across forecast origins k and horizons h. In our implementation, the cross-validation uses an expanding window with an initial calibration window of 10 observations; at each step the window grows by one observation and the next h points are used for validation. We center each horizon-wise residual column, extract all valid consecutive residual blocks of length b, and simulate S future paths by resampling these blocks with replacement and stitching them together along the horizon. This block bootstrap preserves local residual dependence better than horizon-wise independent resampling.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

Algorithm 1 SA-MSCP: Simulation-Enhanced SCP for Aggregated Forecasts

Require: Training series $y_{1:T}$, test start $T+1$, expanding-window CV parameters, simulation count S, horizon H, block size b, quantile levels Q

Ensure: Prediction intervals for monthly $y_{T+1:T+H}$, annual totals, and growth rates

Preprocess $y_{1:T}$.

Fit an algorithm to $y_{1:T}$.

Perform expanding-window CV to collect residual innovations $\{\hat{\varepsilon}_{k,h}\}$.

Center each horizon-wise residual column and extract all valid consecutive residual blocks of length b.

for $s = 1$ **to** S **do**

  Sample residual blocks with replacement until length H, truncate to H, and add them to the forecast path to obtain $\widehat{y}_{s,T+1:T+H}$.

  Aggregate annual totals: $\widehat{Y}_{s,year} = \sum_{m=1}^{12} \widehat{y}_{s,(year,m)}$ for forecast years.

  Compute growth rates $\widehat{G}_{s,year} = (\widehat{Y}_{s,year} - \widehat{Y}_{s,year-1})/\widehat{Y}_{s,year-1}$, with $\widehat{Y}_{s,year-1}$ initialized from last known annual sales in training.

end for

for each target (monthly/annual/growth) and each quantile level $q \in Q$ **do**

  Compute empirical quantiles across $s = 1, \ldots, S$ to form lower/upper bounds.

end for

Return intervals and summary statistics (coverage and width).

In this paper, we apply the method to monthly time series with annual aggregates, but the same approach can be used with other temporal resolutions and aggregation units. Each simulated path yields monthly forecasts, which are aggregated to annual totals $\widehat{Y}_s = \sum_{m=1}^{12} \widehat{y}_{s,m}$ and transformed into growth rates using consecutive annual values. Prediction intervals are then obtained from empirical quantiles of the simulated samples at the desired coverage levels for both pointwise and aggregated targets.

We use block size $b = 12$ for M4, matching its 36-month forecast horizon and preserving annual dependence patterns, and $b = 3$ for the proprietary data. The horizon is only 12 months for the proprietary data, so using $b = 12$ would leave too few distinct bootstrap combinations; $b = 3$ instead preserves within-quarter dependence while retaining sufficient resampling diversity. This simulation-based calibration approximates the distribution of aggregated quantities under temporal dependence, enabling uncertainty quantification beyond pointwise horizons.

We compute empirical coverage rates and interval widths at multiple confidence levels (e.g., 90%, 95%, 99%) across series and forecast scenarios (original, aggregated, growthrate). We assess statistical significance of differences versus a simulated-path baseline using Wilcoxon signed-rank tests.

3. Experiments

We fit Auto-ARIMA models using the ARIMA implementation in the fable package (O'Hara-Wild et al., 2024). Future trajectories are simulated using a modified version of the package's *generate()* function with block-bootstrap innovations drawn from cross-validated residual blocks. For each time series, we generate $S = 10{,}000$ simulated paths.

We evaluate SA-MSCP on M4 monthly sales data for aggregated annual totals (Aggregated Sales) and year-over-year growth rates (Y-o-Y Sales Growth), which are the primary targets of interest, and also report original monthly forecasts (Raw Sales) for reference. We compare against a baseline simulated-path approach (Baseline) (Hyndman & Athanasopoulos, 2021) that uses conditional simulation without cross-validated residual calibration.

A direct split-conformal baseline built on aggregated nonconformity scores is less suitable in our setting because we start from monthly series and evaluate annual aggregates: once the data are aggregated to the yearly level, only a small number of calibration points remain for each series, making direct SCP at the aggregated level statistically inefficient and unstable.

Across all targets and levels, SA-MSCP delivers higher coverage than Baseline with a corresponding increase in interval width, but both methods miss the target coverage level. The coverage and width values are presented in Tables 1 and 2, respectively. In absolute terms, coverage gains range from roughly 6.9 to 17.5 percentage points (Column *Coverage Delta* in Table 3). Relative to Baseline, SA-MSCP improves coverage while widening intervals. The coverage cost summarizes the relative width increase per coverage gain: costs are lowest at the 10% level, increase at 5%, and are highest at 1% (Column *Coverage Cost* in Table 3). Raw Sales exhibits the lowest cost across levels, whereas Y-o-Y Sales Growth is the most costly—especially at 1%. These wider intervals are expected under compounded uncertainty from temporal aggregation.

Real-world monthly sales. We additionally evaluate on 2,000 actual monthly sales series. As with M4, SA-MSCP attains higher coverage with wider intervals at higher confidence. For Raw Sales, SA-MSCP achieves 92.7%, 94.5%, and 96.1% coverage at the 10%, 5%, and 1% levels, respectively. This performance is better than on M4, but it still undercovers except at the 10% miscoverage level. Aggregated

Table 1. Coverage on M4: SA-MSCP vs Baseline

	SA-MSCP			Baseline		
Variant	10%	5%	1%	10%	5%	1%
Raw Sales	88.8%	91.4%	93.9%	75.2%	80.8%	87.0%
Aggregated Sales	83.1%	85.8%	88.9%	65.6%	70.9%	78.4%
Y-o-Y Sales Growth	89.9%	92.0%	94.4%	75.2%	80.5%	87.4%

Table 2. Interval widths on M4: SA-MSCP vs Baseline

	SA-MSCP			Baseline		
Variant	10%	5%	1%	10%	5%	1%
Raw Sales	5.6×10^3	6.4×10^3	7.6×10^3	2.4×10^3	2.8×10^3	3.6×10^3
Aggregated Sales	5.5×10^4	6.4×10^4	7.5×10^4	1.9×10^4	2.2×10^4	2.9×10^4
Y-o-Y Sales Growth	1.6×10^0	2.3×10^0	6.6×10^0	4.5×10^{-1}	5.4×10^{-1}	7.3×10^{-1}

Table 3. Coverage improvements of SA-MSCP over Baseline on M4 data.

Dataset	Level	Coverage Delta	Coverage Cost
Raw Sales	10%	13.6%	10.1
Raw Sales	5%	10.7%	11.8
Raw Sales	1%	6.9%	16.0
Aggregated Sales	10%	17.5%	10.8
Aggregated Sales	5%	15.0%	12.3
Aggregated Sales	1%	10.4%	14.7
Y-o-Y Sales Growth	10%	14.7%	17.2
Y-o-Y Sales Growth	5%	11.5%	28.6
Y-o-Y Sales Growth	1%	7.1%	113.1

Table 4. Coverage on proprietary sales data: SA-MSCP vs Baseline

	SA-MSCP			Baseline		
Variant	10%	5%	1%	10%	5%	1%
Raw Sales	92.7%	94.5%	96.1%	82.8%	87.2%	91.7%
Aggregated Sales	89.3%	91.1%	94.2%	75.3%	80.4%	86.8%
Y-o-Y Sales Growth	89.3%	91.1%	94.2%	75.3%	80.4%	86.8%

Table 5. Interval widths on proprietary sales data: SA-MSCP vs Baseline

	SA-MSCP			Baseline		
Variant	10%	5%	1%	10%	5%	1%
Raw Sales	5.1×10^5	5.9×10^5	7.1×10^5	3.2×10^5	3.8×10^5	4.9×10^5
Aggregated Sales	4.1×10^6	4.8×10^6	6.1×10^6	2.2×10^6	2.7×10^6	3.5×10^6
Y-o-Y Sales Growth	1.4×10^0	1.6×10^0	2.1×10^0	5.9×10^{-1}	7.1×10^{-1}	9.4×10^{-1}

Table 6. Coverage improvements of SA-MSCP over Baseline on proprietary sales data.

Dataset	Level	Coverage Delta	Coverage Cost
Raw Sales	10%	9.8%	6.3
Raw Sales	5%	7.3%	7.5
Raw Sales	1%	4.4%	10.2
Aggregated Sales	10%	14.0%	5.9
Aggregated Sales	5%	10.7%	7.2
Aggregated Sales	1%	7.4%	9.7
Y-o-Y Sales Growth	10%	14.0%	9.4
Y-o-Y Sales Growth	5%	10.7%	12.2
Y-o-Y Sales Growth	1%	7.4%	17.1

Sales and Y-o-Y Sales Growth exhibit slightly lower coverages but the same qualitative pattern. In Table 4, Aggregated Sales and Y-o-Y Sales Growth have identical coverage values because the proprietary test horizon is only 12 months, so annual aggregation produces a single forecast-year total. The corresponding growth rate is then obtained by dividing each simulated annual sum by the same observed annual sum from the previous year. This is a monotone transformation, so if an interval for aggregated sales covers the annual sum, then the corresponding growth-rate interval also covers the associated growth rate; conversely, if the aggregated-sales interval misses the annual sum, then the growth-rate interval also misses the associated growth rate.

Relative to Baseline, SA-MSCP improves coverage by 4.4–14.0 percentage points (Column *Coverage Delta* in Table 6). The coverage cost is smallest for Aggregated Sales at 10% and increases as nominal levels tighten; the largest cost occurs at the 1% level for Y-o-Y Sales Growth, which is lower in magnitude than on M4 but shows the same qualitative trend.

Overall, SA-MSCP achieves materially higher coverage than Baseline across both M4 and proprietary real-world sales data, at the expense of wider intervals. On M4, the width increase per unit coverage gain is smallest for Aggregated at 10% and largest for Growthrate at 1%. On the proprietary dataset, the corresponding extremes are likewise smallest for Aggregated at 10% and largest at 1%, consistent with aggregation effects observed in both settings.

In the absence of exchangeability, conformal prediction methods do not enjoy finite-sample coverage guarantees, and nominal levels should therefore be interpreted as target rather than guaranteed coverage. Consistent with this theoretical limitation, our empirical results show that while SA-MSCP substantially improves coverage relative to simulated-path baselines, achieved coverage remains below nominal levels as shown in Tables 1 and 4. This behavior reflects the intrinsic difficulty of uncertainty quantification in non-stationary and dependent time-series settings, rather than a failure of the proposed method.

In practice, users requiring closer alignment between nominal and achieved coverage may apply post-hoc calibration strategies, such as increasing the nominal coverage level, or applying online adaptation schemes introduced in (Gibbs & Candès, 2021). Investigating principled calibration schemes that account for temporal dependence is an important direction for future work.

Significance testing. Wilcoxon signed-rank tests comparing SA-MSCP vs Baseline yield p-values $\ll 0.001$ for coverage and width across Aggregated and Growthrate targets at 10%, 5%, and 1% levels. Differences are statistically significant.

4. Conclusion

We propose SA-MSCP, a simulation-enhanced multi-step split conformal method that combines expanding-window cross-validation, block-bootstrap future-path simulation, and calibration on aggregated trajectories. SA-MSCP improves empirical coverage for monthly, aggregated annual, and growth-rate targets, at the cost of wider intervals, reflecting a fundamental trade-off in conformal prediction under dependence and aggregation. Residuals across forecast origins and horizons are not strictly exchangeable; expanding-window cross-validation partially accounts for temporal dependence by pooling residuals over time, while the block bootstrap preserves local cross-horizon dependence in simulated paths. Sequential CP methods (e.g., SPCI, PID) target pointwise horizons and long-run coverage, making extension to aggregated quantities nontrivial. Copula-based calibration could model cross-horizon dependence for sums and growth rates, but requires multiple time series, whereas our setting focuses on a single series (Sun & Yu, 2022).

Overall, SA-MSCP provides a practical and scalable approach to uncertainty quantification for aggregated forecasting tasks.

References

Angelopoulos, A. N., Candès, E. J., and Tibshirani, R. J. Conformal pid control for time series prediction. *Advances in Neural Information Processing Systems*, 36: 23047–23074, 2023.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. Conformal prediction beyond exchangeability. *The Annals of Statistics*, 51(2):816–845, 2023.

Gibbs, I. and Candès, E. Adaptive conformal inference under distribution shift. In *Advances in Neural Information Processing Systems*, volume 34, pp. 1660–1672, 2021.

Hyndman, R. J. and Athanasopoulos, G. *Forecasting: principles and practice*. OTexts: Melbourne, Australia. OTexts.com/fpp3, 3rd edition, 2021.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111, 2018.

Luo, R. and Zhou, Z. Conformalized interval arithmetic with symmetric calibration. In *Proceedings of AAAI-25*, 2025.

O'Hara-Wild, M., Hyndman, R. J., and Wang, E. *fable: Forecasting Models for Tidy Time Series*, 2024. URL https://fable.tidyverts.org. R package version 0.4.0.

Principato, G., Stoltz, G., Amara-Ouali, Y., Goude, Y., Hamrouche, B., and Poggi, J.-M. Conformal prediction for hierarchical data. *arXiv preprint arXiv:2411.13479*, 2025.

Shafer, G. and Vovk, V. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9:371–421, 2008.

Stankevičiūtė, K., Alaa, A., and van der Schaar, M. Conformal time-series forecasting. In *Advances in Neural Information Processing Systems*, volume 34, pp. 6216–6228, 2021.

Sun, S. and Yu, R. Copula conformal prediction for multi-step time series forecasting. *arXiv preprint arXiv:2212.03281*, 2022.

Tibshirani, R. J., Foygel Barber, R., Candes, E., and Ramdas, A. Conformal prediction under covariate shift. *Advances in neural information processing systems*, 32, 2019.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, 1st edition, 2005.

Wang, X. and Hyndman, R. J. Online conformal inference for multi-step time series forecasting. *arXiv preprint arXiv:2410.13115*, 2024.

Xu, C. and Xie, Y. Conformal prediction interval for dynamic time-series. In *International Conference on Machine Learning*, pp. 11559–11569, 2021.

Xu, C. and Xie, Y. Sequential predictive conformal inference for time series. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 38707–38727, 2023.

Zou, X. and Liu, W. Coverage-guaranteed prediction sets for out-of-distribution data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 17263–17270, 2024.