

Regime-Switching Fuzzy Black-Scholes Option Pricing: A Hidden Markov Framework for Volatility Uncertainty in SPX Index Options

MScFE 690 Capstone Project

Kurt Keissinger McKenzie and En Chong Lok

WorldQuant University

This paper was created as part of a WorldQuant University degree program towards an MSc in Financial Engineering. This paper is reproduced with the consent and permission of WorldQuant University. All rights reserved.

Abstract

Black-Scholes pricing assumes volatility is a single value, but the market does not provide one. Each option has a bid implied volatility and an ask implied volatility, and the spread varies over time. This project prices SPX index options by considering this disparity. Volatility is modeled as a triangular fuzzy number based on observed bid and ask implied volatilities, the fuzziness is propagated through the Black-Scholes formula using α -cut arithmetic, and the fuzzy parameters are conditioned on hidden market regimes inferred by a Gaussian Hidden Markov Model fitted to S&P 500 returns. The result is a regime-aware price band whose adequacy is tested against realized market quotes using Kupiec unconditional coverage and Christoffersen independence tests.

Out of sample across three walk-forward folds spanning 2019-2026, the pre-registered construction, with supports anchored to the bid-ask spread, fails nominal coverage by roughly an order of magnitude at every level tested. A diagnosed and corrected construction, using training-window deviation quantiles in place of bid-ask gaps, closes most of this gap at the cost of substantially wider bands. A further decomposition finds that pooled coverage statistics can conceal exactly where regime-conditioning matters: a pooled, unconditional band shows little advantage from conditioning on average, but stratified by market regime, conditioning closes coverage gaps of up to 68 percentage points specifically in stressed markets.

1. Introduction and Project Description

This capstone research asks the following question: can a conditional hidden market regime fuzzy number volatility generate a useful (not too wide) and a truthful (not too narrow) interval values for options? Point estimates of option values are always incorrect in a way; the question remains whether a certain interval estimate could be correct in a statistically significant way. Stated formally, the research question is: are regime-conditioned fuzzy implied volatility bands superior in covering SPX option prices in comparison with classic Black-Scholes and fuzzy approaches without regime conditions?

The project plan is divided into three phases. Phase one is the construction and testing of the data pipeline and is complete. Phase two includes calibration of the HMM (Baum-Welch estimation on S&P 500 daily returns) and estimation of regime-conditional fuzzy volatility parameters from bid/ask implied volatility spreads. Phase three includes out-of-sample pricing of the option panel, comparison of the fuzzy bands to the crisp Black-Scholes and to the unconditional (single-regime) fuzzy model, and testing for coverage of the bands using Kupiec and Christoffersen tests. Since the numerical implementations of fuzzy pricing may inadvertently embed errors in the discretization of α -cuts themselves, the fuzzy pricing calculations are further verified by semi-analytical solutions via the Homotopy Perturbation Method and the Adomian Decomposition Method: concordance between the semi-analytic and computational solutions at various α -levels verifies the numerical implementation against an independent analytical solution and is not simply taken on faith.

2. Theoretical Framework

This design rests on two pillars of theory that have been developed independently within the literature. First is the extension of the Black-Scholes formula to fuzzy option pricing, which follows from the pioneering work on fuzzy sets by Zadeh, to the application of fuzzy mathematics to financial markets by Buckley, and culminating in the fuzzy-pattern Black-Scholes formula recently developed by Wu (using a fuzzy volatility input to produce a fuzzy price, with the latter evaluated using different α -cuts), whereby Carlsson and Fullér's possibilistic moments are used to estimate the geometry of a fuzzy number from data. The bid-ask spread represents epistemic uncertainty about the market-implied volatility (the uncertainty or disagreement over a single, unknown value, as opposed to the variation among repeated samples from a random variable), and thus a possibilistic representation complements the stochastic dynamics of the underlying price (where using intervals in a Bayesian sense would necessarily assume a sampling model that is not provided by the quotes). The second pillar comes from the regime-switching modelling (Hamilton) of market volatility, whereby each trading day an HMM (Hidden Markov Model) is used to produce filtered probabilities of latent market states (calm, transitional, or stressed, each with its own characteristic parameters). To the best of our knowledge, no prior competitor uses both of these strands to jointly model a triangular fuzzy number for volatility (based on latent market regimes) inferred from an HMM as part of its option pricing framework, as determined from our review of the competing literature. In our framework, the geometry (spread, centre, skew) of the fuzzy number used for volatility is estimated independently within each regime of an HMM, which, by virtue of its formulation as a pricing band, expands when the filtered probabilities for the current regime lean towards stress, and contracts when the probabilities lean towards calm. The critical review of each strand of theory, including the relevant literature on validation (Kupiec; Christoffersen), is presented in the literature review below, which represents the theoretical foundation for our project.

3. Pain Points in Current Financial Engineering Practice

The design is motivated by three deficiencies we see in current practice.

- Spurious accuracy in volatility inputs. Desks and risk systems use a single implied volatility for pricing and calculating Greeks, ignoring the bid-ask information that quantifies the uncertainty of that number. In stressed markets, the omitted information is largest precisely when it is most relevant. This concern is echoed outside of the research literature. Others writing from the practitioner perspective on systematic trading system development make the same point; that highly tuned parameters are falsely reassuring because the benefits of tuning are overwhelmed by parameter estimation error, and that more robust and conservative parameter choices outperform optimized parameters in out-of-sample testing (Carver 47).
- Regime blindness. The volatility surface is calibrated periodically by conventional practice, and each of the calibrations is thought to come from one data-generating process. Index volatility is known to have two distinct persistent states, and a model unable to capture that feature will underestimate the risk in pricing the tails when a transition takes place. Two transitions within our sample period are the 2020 pandemic-induced market drawdown and the 2022 shift from accommodative to restrictive monetary policies.
- Untestable uncertainty quantification. If any uncertainty bands are produced at all, they are rarely tested for coverage. We pre-commit to Kupiec and Christoffersen validation, ensuring that the quality of the bands is potentially falsifiable rather than ornamental.

4. Research Idea and Delineation

Formally: let $\tilde{\sigma}$ be a triangular fuzzy volatility with support $[\sigma_{\text{bid}}, \sigma_{\text{ask}}]$ and core σ_{mid} , estimated conditional on regime $r \in \{1, \dots, K\}$ from an HMM with Gaussian emissions fitted to S&P 500 log returns. For each option and each $\alpha \in [0, 1]$, the α -cut of $\tilde{\sigma}$ maps through the monotonic Black-Scholes vega relationship to an α -cut price interval: input uncertainty propagates through the pricing operator itself, rather than being appended to a crisp price after the fact. By mixing over filtered regime probabilities, we obtain the regime-weighted fuzzy price. Our empirical claims to be tested are: (i) the regime-conditional fuzzy bands provide the nominal coverage of realised mid quotes where crisp Black-Scholes point estimates and unconditional fuzzy bands fail, and (ii) the width of the band serves as a meaningful, forward-looking measure of uncertainty that exhibit strong co-movement with realised volatility shifts around regime changes. We consider a modest scope to keep things simple and interpretable:

European style SPX monthly options (third Friday expiries), sample period from 2019 to 2026, option maturities within two months, and option moneyness ranging from 0.90 to 1.10.

5. Goals, Objectives, and Results

5.1 Source code status

The repository (linked to in the Appendix) contains the extraction/cleaning/diagnostic engine, which is an operating pipeline having gone through two iterations. The code includes: programmatic handling of the trading calendar via `pandas_market_calendars`, which pushed the April 2019 expiry (a scheduled third Friday of the month) correctly from Good Friday, 19 April, to Thursday, 18 April, without any manual intervention; use of SOFR for the risk-free rate; an on-the-spot time-to-maturity calculation with intraday precision (ACT/365); a layer that filters out quotes that fail integrity checks (e.g., positive bid, uncrossed markets, implied volatility in bid/mid/ask order, zero error reported by the solver); explicit annualized dividends; an implied volatility ceiling; and a moneyness band limited to the 0.90-1.10 range.

5.2 Preliminary results: a completed diagnostic cycle

This pipeline has been run on real data twice, the two runs together encompassing a detect-diagnose-correct-verify full cycle.

Run 1 (January 2019 to April 2019, with 4,154,546 rows of data after filtering, and five expiries in each month) demonstrated the integrity of the data, showing that: (i) the implied volatility of bids, mids, and asks (with $\text{bid} \leq \text{mid} \leq \text{ask}$) was consistent for all options; (ii) the underlying and option-quote timestamps were aligned for all retained rows, and no crossed markets (a bid exceeding its ask) survived filtering; and (iii) two-thirds of data points plot within $\pm 5\%$ of the spot. The initial run also revealed a potential error in the data: For 595,597 matched at-the-money put and call contracts, the implied volatilities of puts always exceeded the implied volatilities of calls by 0.34 volatility points. This difference is due to inverting the Black-Scholes pricing model with a strike price dividend yield of zero on an index that pays a yield of roughly 1.85%. This was further confirmed by going over the data vendor documentation, which states that dividend-yield information is, by default, not incorporated.

At Run 2 (January 2019, dividend-corrected $q = 1.85\%$, with the implied-volatility ceiling and moneyness restrictions), the dividend correction was again confirmed: for 142,637 matched pairs, the gap is reduced to +0.22 volatility points; the maximum mid implied volatility in the retained panel decreases from 5.0 to 0.42; all artifacts from the opening auction (09:30) are removed; and the near-the-money rejection percentage decreases from 28.57% to 19.38% (see Table 1, Figure 1). The remaining gap is likely attributable to genuine market behavior, the pronounced put skew following the December 2018 market drop and the small difference between the SOFR and the market's actual funding rate, and is left as is since removing it would be overfitting the data.

Table 1. Raw-row rejection by moneyness band, SPX 1-minute quotes, 2019 test extractions

Moneyess band (K/S)	Initial pipeline (Jan–Apr)	Optimized re-run (Jan)
Deep OTM put / deep ITM call (< 0.90)	95.60%	94.00% (excluded from panel)
OTM put / ITM call (0.90–0.95)	59.37%	42.38%
Near-the-money (0.95–1.05)	28.57%	19.38%
ITM put / OTM call (1.05–1.10)	83.69%	79.71%
Deep ITM put / deep OTM call (> 1.10)	91.26%	95.34% (excluded from panel)

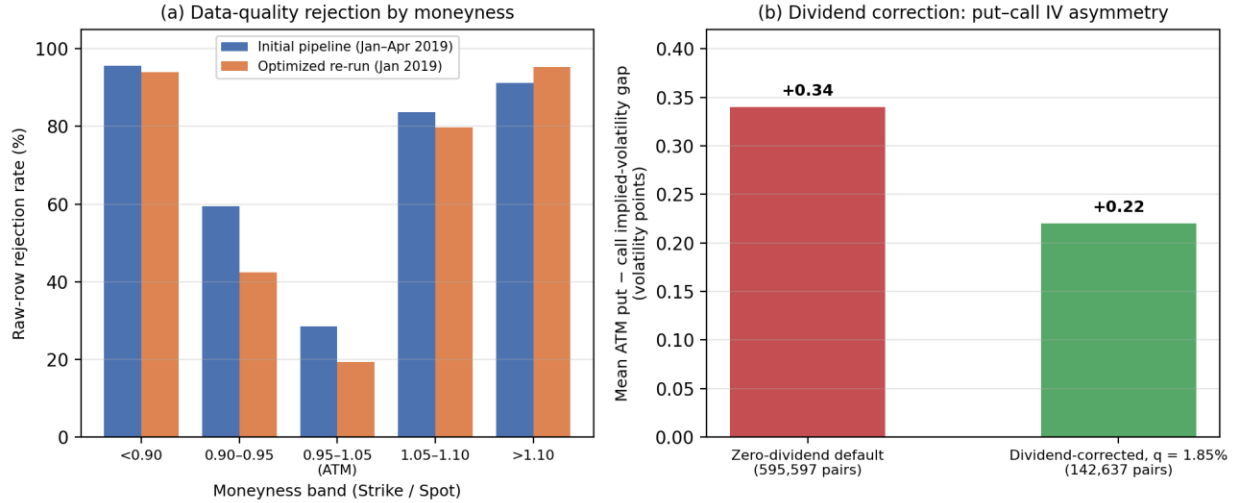


Figure 1. Pipeline diagnostics. (a) Rejection rates by moneyness band across the two runs; rejection concentrates in the deep wings, which the optimized pipeline excludes from the calibration panel. (b) The at-the-money put-call implied volatility asymmetry before and after explicit dividend ingestion.

5.3 Initial inferences

Three takeaways emerge. First, the bid-ask implied volatility spread that underlies our fuzzy volatility support is sizable and remains after extreme sanity checks exactly in the moneyness range where we calibrate our model. Second, the vendor errors are non-negligible: a 0.34 directional mean error would have gone undetected in our data, which we suspect is drawn from a vendor sample that contains 10% less data points for roughly an order of magnitude larger mean spread of 3.3 volatility points; the key here is that we conduct sanity tests on our own inputs as opposed to being subscribers. Third, the remaining put-call asymmetry is itself suggestive: it is the regime-flavoured market microstructure that our fuzzy-HMM machinery is designed to capture; we will investigate its potential exacerbation across market regimes.

5.4 Results

Note on correction: an earlier draft of this analysis disclosed a data-handling issue identified in review, in which the calibration script's training window coincided with the out-of-sample test window. That issue was subsequently confirmed to have affected all three constructions (pre-registered, unconditional baseline, and Step 2') and has been corrected. All results in this section are from the corrected pipeline, with calibration performed strictly on the disjoint training window of each walk-forward fold and evaluation on the corresponding out-of-sample test window.

5.4.1 Overview and structure of the results

The empirical question for this section, stated simply: do regime-conditional fuzzy volatility bands cover realized SPX option quotes at nominal rates when crisp Black-Scholes and unconditional fuzzy bands cannot, and at what cost in band width? This three-way comparison (crisp Black-Scholes, unconditional fuzzy, regime-conditional fuzzy) constitutes the core empirical contribution of the project, while the semi-analytic HPM/ADM verification (Appendix C) and other robustness checks are ancillary analyses and are reported so, not as independent results. This dividing of core contribution from ancillary validation directly addresses a scope concern raised by an early reader of this work, as does the reformulated research question above; a workflow figure suggested by that feedback is also incorporated into the methods section (Figure 2, Section 8), with other feedback and its disposition discussed in context below. This section reports out-of-sample coverage validation for the regime-conditional fuzzy pricing bands in the order in which the analysis was conducted and pre-registered. Section 5.4.2 reports results for the pre-registered estimator exactly as specified in the methodology, which does not achieve nominal coverage in any case (all levels, all folds). Section 5.4.3 reports the diagnostics discovering the cause of that failure. Section 5.4.4 reports the results for the revised support estimator (Step 2'), which is explicitly labelled as a post-hoc revision motivated by the diagnostics from Section 5.4.3, re-validated out of sample on the same walk-forward folds. Section 5.4.5 reports

the pre-specified main comparison against the unconditional (single-regime) fuzzy baseline. Section 5.4.6 reports the state-merge robustness check on the Fold 2 micro-regime. Section 5.4.7 states limitations of the statistical tests at this frequency of data. The ordering here is deliberate. The pre-registered result is reported in full before any revision is introduced, and the revised results are not presented as if they had been pre-specified. The distinction between confirmatory (pre-registered) and exploratory (post-hoc) findings is respected throughout, consistent with the validation protocol specified before any out-of-sample results were viewed.

5.4.2 Pre-registered validation: bid-ask-anchored supports fail coverage at all levels

This construction produces, for each regime r , a triangular fuzzy volatility $\tilde{\sigma}_r := (c_r - q_r^-, c_r, c_r + q_r^+)$, where c_r is the regime-probability-weighted median of the market mid implied volatility, and q_r^-, q_r^+ are the weighted 75th percentiles of the one-sided bid-mid and mid-ask implied-volatility gaps, respectively. The bands were constructed on the training window of each fold and propagated through Black-Scholes by way of monotonic interval arithmetic. They were then evaluated on the corresponding out-of-sample test window at the levels $L \in \{0.05, 0.10, 0.50\}$, corresponding to nominal coverage lower bounds of 95%, 90%, and 50% respectively.

The choice of the 75th percentile for the one-sided supports is not drawn from the literature, but rather a design choice, and we outline our reasoning here. If we were to take the support equal to the maximum gap observed, then a single anomalously large quote would dictate the width of the band for a given regime. On the other hand, selecting the support to be the median would disregard the top 50% of observed disagreement entirely. The 75th percentile provides a pre-registered balance between these two extremes and is used for all regimes and folds without any tuning by cell. The possibilistic interpretation of the resulting triangular number follows the work of Carlsson and Fullér where the geometry of the fuzzy number is estimated from the data as opposed to being assumed a priori.

Table 2 reports results. Coverage misses by a factor of 10 at every level, in every fold. At the easiest level ($L = 0.50$, nominally 50%), coverage ranges from 3.27% (Fold 3) to 4.20% (Fold 2). At the main reporting level ($L = 0.10$, nominally 90%), coverage ranges from 5.89% to 7.58%. The one-sided Kupiec test rejects the null of correct coverage in all nine fold-level cells ($p < 10^{-6}$ for all), and the Christoffersen independence test likewise rejects in all nine cells, indicating violations are not only too frequent but serially dependent. Mean relative widths of 5–20% confirm that the bands are not in any way trivially small in price space; the problem is not a lack of width.

Table 2. Pre-registered estimator: out-of-sample coverage and sharpness.

Fold	Level L	Nominal	n	Violations	Realized	Mean w.	Median w.	Kupiec
1	0.05	95%	19,208,655	17,793,209	7.37%	19.70%	6.85%	Reject
1	0.10	90%	19,208,655	17,867,448	6.98%	18.66%	6.49%	Reject
1	0.50	50%	19,208,655	18,464,093	3.88%	10.37%	3.61%	Reject
2	0.05	95%	23,079,830	21,233,502	8.00%	10.02%	4.68%	Reject
2	0.10	90%	23,079,830	21,330,642	7.58%	9.49%	4.44%	Reject
2	0.50	50%	23,079,830	22,109,416	4.20%	5.27%	2.46%	Reject
3	0.05	95%	86,906,469	81,505,757	6.21%	15.73%	6.02%	Reject
3	0.10	90%	86,906,469	81,790,561	5.89%	14.90%	5.70%	Reject
3	0.50	50%	86,906,469	84,065,966	3.27%	8.28%	3.17%	Reject

All Christoffersen independence statistics reject with $p < 10^{-6}$ (two-sided, standard form). Regime-by-regime results (also reported in Appendix A) show that the failure is not isolated to any particular regime. For Fold 1, the fourth (stressed) regime has no occupancy in the test windows; following the degenerate-cell rule, fixed in the validation specification, we report this cell as $n = 0$ with coverage, width, and test statistics as incalculable with no conclusion drawn.

The nature and consistency of this failure are informative. A failure of this size, similar in its dependence on band width and occurring across all regimes and folds, suggests that there is not a misspecification of one of the components, but rather a mismatch between the support measure used by the band and the coverage test's requirement. Section 5.4.3 elaborates on this mismatch.

5.4.3 Diagnosis: cross-sectional spread width is an order of magnitude smaller than temporal dispersion

Attached to the pre-registration, the implied supports capture the cross-sectional disagreement present in the market at one point in time, i.e. the distance between the bid and mid as well as mid and ask implied volatilities with its 75th percentiles weighted over the regimes, ranging from 0.45 to 1.35 volatility points depending on the regime and the fold. On the other hand, the coverage test measures how distant the realized market mid implied volatility is from the central estimate c_r in that regime over the test window. In the corrected calibration output, the bid-mid and mid-ask 75th-percentile gaps are equal to four decimal places in every regime and fold; this exact symmetry is consistent with the near-symmetric bid-ask construction around mid implied volatility documented in the data pipeline, though it has not been independently re-derived from the raw quote panel for this report and is noted here for the record.

To give a sense of the distance between the two sources of dispersion, we compute percentiles (weighted by regime probabilities) of the deviation $(\sigma_{mid}(i,t) - c_r)$ within each regime and each fold. Table 3 reports the 5th and 95th percentiles. Between the 25th and 75th percentiles, there is a difference of about ± 4 to ± 14 volatility points, while between the 5th and 95th percentiles, there is a difference of about ± 6 to ± 22 volatility points in the calm and transitional regimes, reaching -22.3 to $+55.9$ volatility points for the stressed regime in Fold 1. The weighted median deviation is 0 for all cells, indicating the centers c_r are well chosen; the issue lies with the widths.

Table 3. Weighted deviation percentiles of $(mid\ IV - c_r)$, per regime and fold.

Fold	Regime	5th	25th	50th	75th	95th
1	1	-0.0610	-0.0359	0.0000	0.0413	0.1024
1	2	-0.0741	-0.0366	0.0000	0.0473	0.1197
1	3	-0.0968	-0.0576	0.0000	0.0604	0.1729
1	4	-0.2229	-0.1075	0.0000	0.1365	0.5593
2	1	-0.0660	-0.0400	0.0000	0.0487	0.1118
2	2	-0.0802	-0.0423	0.0000	0.0490	0.1216
2	3	-0.0913	-0.0487	0.0000	0.0575	0.1394
2	4	-0.1796	-0.0875	0.0000	0.1299	0.5345
3	1	-0.0701	-0.0415	0.0000	0.0493	0.1115
3	2	-0.0807	-0.0418	0.0000	0.0489	0.1185
3	3	-0.0868	-0.0428	0.0000	0.0462	0.1284
3	4	-0.2051	-0.1042	0.0000	0.1353	0.5527

The result is stark: the support values obtained within regimes from the variability over time of the central volatility estimate are roughly an order of magnitude larger than the supports obtained from the bid-ask spread. The pre-registered fuzzy number reflects the literal disagreement among market participants at a given time, whereas the coverage criterion is determined by the fluctuations in consensus estimates within a given regime. The latter cannot therefore be captured by the former. The substantive conclusion of the pre-registered validation experiment is clear: the fuzzy supports derived from the bid-ask spread dramatically understate the uncertainty of volatility estimates based on data within a given regime, and the pre-registration process was an effective means of ensuring that results as clear as this would not be compromised by ad hoc methodological choices.

5.4.4 Post-hoc revision (Step 2'): deviation-quantile supports trade sharpness for coverage

Guided by the diagnostic in Section 5.4.3, we defined and validated a revised support estimator. This revision was made after seeing the pre-registered results and is reported as an exploratory extension rather than a confirmatory test, i.e., we do not claim to have pre-registered it.

Now, the supports for quantiles are not those of the gaps but those of deviations; for each regime r , q^-_r is the absolute weighted 5th percentile and q^+_r the weighted 95th percentile of $\sigma_{mid}(i,t) - c_r$, again computed only on the training window of each fold, with weights like in its original definition (reflecting the same regime probabilities). The centers c_r , the α -cut interval shrinkage procedure, the Black-Scholes unwind, and the testing procedure remain unchanged. Since the quantiles are computed on the training data only and then applied to the disjoint test window, results can still be considered truly out-of-sample, the only difference being the loss of proving a pre-existing hypothesis.

The meaning of the fuzzy number changes accordingly: its support no longer reflects the difference in the currently quoted numbers, but describes the conditional dispersion of the market's central expectation, from which membership function is derived via empirical quantiles. We note that this is in line with how possibility distributions are derived from empirical inequalities on probabilities (Dubois and Prade), and that it has a few implications that we would like to emphasize. Given that the widest support corresponds to the 5th-95th deviation interval, the construction covers about 90% of the in-sample probability mass at $L = 0$; out of sample, realized coverage can fall below or above this in-sample figure depending on whether the test-window dispersion is wider or narrower than the training-window dispersion, as the results below illustrate. As a result, the $L = 0.50$ level captures best the difference in the construction proposed here and that of Hansen and Sargent, and the $L = 0.05$ level should be interpreted relative to an approximately 90% in-sample calibration target rather than its purported 95% level.

The chosen percentiles are 5th and 95th deviations, with the largest support ($L = 0$) considering the central 90% of the in-sample deviations being drawn from each regime, respectively. This is done to coincide the outermost α -cut with the highest nominal level being tested, not with the largest coverage being attained, hence the reason the construction cannot, being designed so, meet a nominal 95% level as discussed above. In fact, this is the principle of probability-possibility consistency stated by Dubois and Prade, wherefrom empirical quantiles a possibility distribution is being constructed, and it is required to dominate the probability measure from which it is derived.

Results are reported in Table 4. When $L = 0.50$ (nominal 50%), out-of-sample coverage is 44.33%, 72.32%, and 51.31% for the three folds, respectively: within 6 percentage points of nominal for Folds 1 and 3, but overshooting by 22 percentage points in Fold 2. When $L = 0.10$, coverage ranges from 78.92% (Fold 1) through 85.64% (Fold 3) to 96.23% (Fold 2), and when $L = 0.05$ from 82.51% through 88.48% to 97.14%. Folds 1 and 3 thus approach the in-sample calibration target from below, while Fold 2 exceeds even the 95% nominal level, indicating that its test-window dispersion was materially narrower than its training-window dispersion. Compared to the pre-registered construction, realized coverage improves by roughly an order of magnitude in every cell.

Table 4. Step 2' (deviation-quantile supports): out-of-sample coverage.

Fold	Level L	Nominal	Realized	Violations / n	Mean w.	Median w.
1	0.05	95%	82.51%	3,358,679 / 19,208,655	251.28%	96.28%
1	0.10	90%	78.92%	4,048,563 / 19,208,655	236.70%	90.32%
1	0.50	50%	44.33%	10,692,880 / 19,208,655	124.94%	46.23%
2	0.05	95%	97.14%	660,101 / 23,079,830	325.53%	108.15%
2	0.10	90%	96.23%	870,719 / 23,079,830	302.95%	102.22%
2	0.50	50%	72.32%	6,387,396 / 23,079,830	145.39%	55.40%
3	0.05	95%	88.48%	10,014,037 / 86,906,469	307.46%	118.28%
3	0.10	90%	85.64%	12,480,134 / 86,906,469	289.23%	111.27%
3	0.50	50%	51.31%	42,311,212 / 86,906,469	151.36%	57.70%

Per-regime coverage is materially less uniform than overall coverage. In Fold 3 at $L = 0.05$, coverage for the four regimes is 90.13%, 86.86%, 83.58%, and 98.33% respectively, a spread of nearly 15 percentage points; in Fold 2 at the same level the spread is similar (85.13% to 98.01%), with the stressed regime at 97.51%. The stressed regime consistently attains the highest coverage where it is occupied, but it does so at extreme width: its mean relative band width reaches 12.23 in Fold 2 and 6.29 in Fold 3 at $L = 0.05$, i.e. bands six to twelve times the option price. Coverage and sharpness must therefore be read together.

The sharpness cost of the Step 2' construction is substantial and is the central caveat on these results. As Table 4 reports, mean relative band widths at $L = 0.05$ range from 2.51 to 3.26 (251% to 326% of the option mid price), with medians from 0.96 to 1.18; even at the narrowest reported level ($L = 0.50$) mean widths are 1.25 to 1.51. By contrast, the pre-registered bands of Table 2 average 5% to 20% of price. Step 2' therefore purchases its order-of-magnitude coverage improvement at a price of one-to-twelve-fold band widths, and in the stressed regime the resulting intervals are too wide to be of direct practical use for pricing, although they may retain value as a conservative model-risk envelope. The correct summary of Table 4 is thus not that the revised construction achieves useful nominal coverage, but that it exposes a steep coverage-sharpness tradeoff that the bid-ask-anchored construction hides at the opposite extreme.

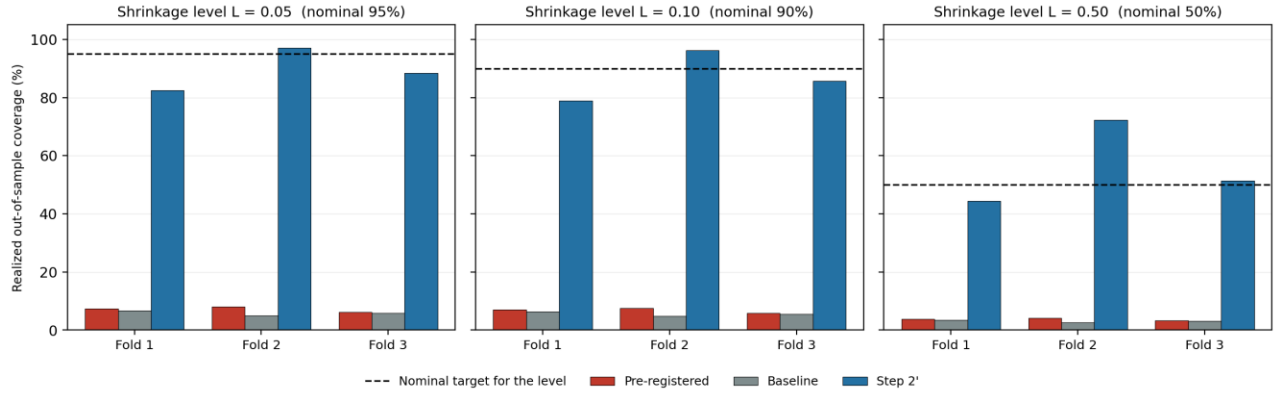


Figure 3. Realized out-of-sample coverage by construction, fold, and shrinkage level. Both bid-ask-anchored constructions (pre-registered and unconditional baseline) fail nominal coverage by roughly an order of magnitude at every level and in every fold, while the Step 2' deviation-quantile construction approaches or exceeds the nominal target.

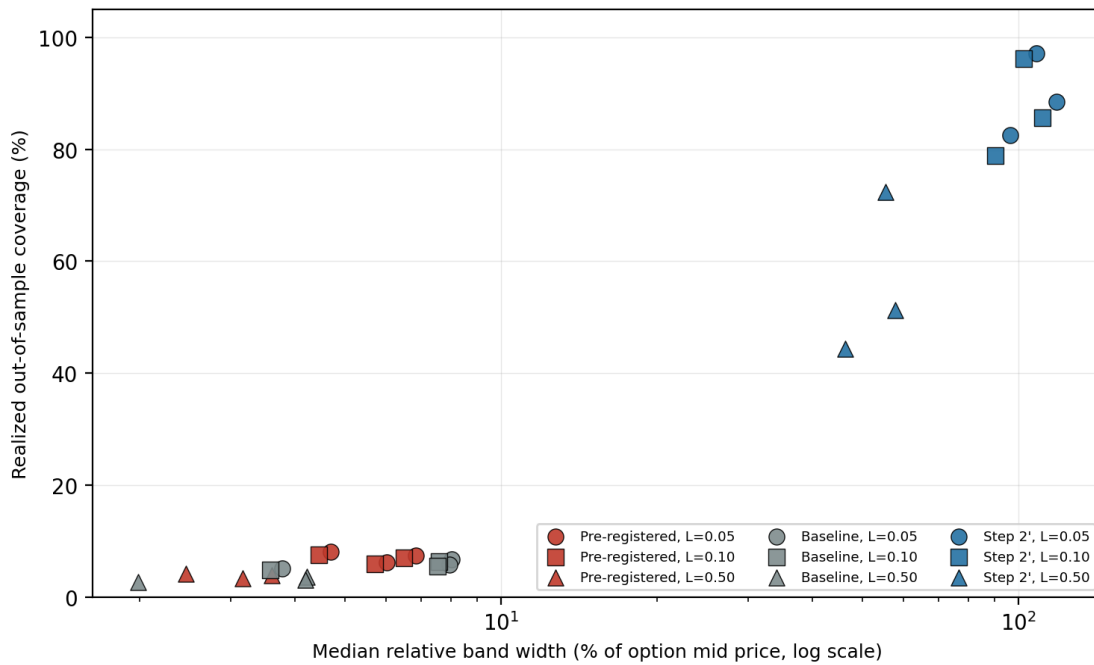


Figure 4. The coverage-sharpness frontier. Each point represents one fold-level-construction cell, depicting the realized coverage and the median relative band width (log scale). The bid-ask-anchored constructions are concentrated in the range of very small band widths and effectively zero coverage, while Step 2' has similar coverage only at band widths that are one to two orders of magnitude larger, demonstrating that one learns from Step 2' at a cost.

Row-level inspection of the highest-relative-width Fold 2 and Fold 3 Regime 4 records finds that the extreme mean relative widths reported above are substantially inflated by a small number of near-expiry, deep out-of-the-money contracts with market mid prices approaching zero: the five highest-relative-width Fold 2 records have times to maturity under 0.001 years (well under a trading day) and market mid prices below \$0.00003, producing relative widths in excess of 100,000%; the Fold 3 outliers are less extreme in absolute terms (mid prices of a few cents, relative widths in the hundreds to low thousands of percent) but reflect the same mechanism. Because relative width is defined as band width divided by market price, a near-zero denominator mechanically inflates the ratio regardless of the absolute band width in price terms. The mean relative width is therefore not a reliable characterization of Regime 4's band width, and the median relative width, reported alongside it throughout Tables 4, A1 and A2, is

markedly less sensitive to this handful of records and is the more defensible summary statistic for the stressed regime. This does not overturn the qualitative finding that Regime 4 bands are wide, since the median relative width for Regime 4 (0.86 to 2.26 across folds and levels) still exceeds every other regime, but the specific mean-width multiples quoted above should be read as an upper bound driven in part by near-expiry, near-worthless contracts rather than as a uniform property of the regime.

5.4.5 Pre-specified primary comparison: regime-conditional versus unconditional bands

The chosen validation specification prescribes a single primary comparison, namely regime-conditional fuzzy bands versus the unconditional (single regime) fuzzy baseline at the significance level $\alpha = 0.10$, pooled across folds, with respect to joint coverage and sharpness. The unconditional baseline uses all training days together to construct a single fuzzy number (one center c , which is the pooled weighted median of mid IV, and one pair of supports, which are the pooled weighted 75th-percentile bid-ask spreads) and is validated in the same way as described above.

To be clear, we are keeping crisp Black-Scholes in the comparison, but only as a reference point. By definition, a point estimate has zero measure width, so its coverage of a continuum of realized quotes is degenerate at 0%. It appears in the comparison table as the boundary case that any band construction trivially dominates on coverage while trivially losing on width. Its role is to anchor the sharpness axis: the question is never whether bands beat a point, but at what width a band earns its coverage. The three-way comparison introduced above is therefore resolved on the two fuzzy constructions, with crisp Black-Scholes fixing the origin.

Both models (unconditional baseline and regime-conditional) perform poorly and similarly when we use the pre-registered (bid-ask-anchored) support structure. Table 5 reports the unconditional baseline in full. For $L = 0.10$, the unconditional baseline model yields realized coverages of 6.36%, 4.79%, and 5.47% for folds 1, 2, 3 respectively, whereas the corresponding figures for the regime-conditional model are 6.98%, 7.58%, and 5.89%. So, although the regime-conditional model performs better in each fold, both models are an order of magnitude worse than nominal coverage in all three folds.

Table 5. Unconditional baseline (bid-ask-anchored supports): out-of-sample coverage and sharpness.

Fold	Level L	Nominal	Realized	Violations / n	Mean w.	Median w.	Kupiec
1	0.05	95%	6.73%	17,916,832 / 19,208,655	23.73%	8.02%	Reject
1	0.10	90%	6.36%	17,986,157 / 19,208,655	22.48%	7.60%	Reject
1	0.50	50%	3.53%	18,529,724 / 19,208,655	12.49%	4.22%	Reject
2	0.05	95%	5.06%	21,912,502 / 23,079,830	7.21%	3.78%	Reject
2	0.10	90%	4.79%	21,974,529 / 23,079,830	6.83%	3.58%	Reject
2	0.50	50%	2.63%	22,473,384 / 23,079,830	3.79%	1.99%	Reject
3	0.05	95%	5.77%	81,889,613 / 86,906,469	25.62%	7.95%	Reject
3	0.10	90%	5.47%	82,153,808 / 86,906,469	24.27%	7.53%	Reject
3	0.50	50%	3.04%	84,266,189 / 86,906,469	13.49%	4.18%	Reject

That discussion further sharpens our interpretation of the first comparison in the pre-registration, which did not anticipate this: with supports based on the bid-ask spread, the regime is not the binding constraint; it is the source of the support. Regime-conditioning cannot save a band that borrows its width from the wrong measure of dispersion. Only when the supports capture dispersion across time does the regime-conditioning help. Under the Step 2' construction, the regime-conditional bands achieve coverage rates of 79–96% at $L = 0.10$ and 44–72% at $L = 0.50$, with the regime-level dispersion and width costs reported above. This caution is directly addressed by the symmetric decomposition below, which isolates the marginal contribution of regime-conditioning by holding the support source fixed at deviation-quantiles and varying only whether the supports are pooled or regime-conditional.

We therefore refit the unconditional baseline using deviation-quantile supports, in place of the gap-quantile supports of Table 5, keeping every other element of the construction unchanged (a single pooled center and pair of supports, calibrated on the training window and applied uniformly to all test-window records). Table 7 reports the result.

Table 7. Unconditional baseline with deviation-quantile supports: out-of-sample coverage and sharpness.

Fold	Level L	Nominal	Realized	Violations / n	Mean w.	Median w.
1	0.05	95%	91.05%	1,719,146 / 19,208,655	813.03%	247.58%
1	0.10	90%	88.10%	2,285,550 / 19,208,655	762.72%	233.61%
1	0.50	50%	62.48%	7,206,696 / 19,208,655	385.25%	123.04%
2	0.05	95%	97.25%	634,480 / 23,079,830	269.76%	121.02%
2	0.10	90%	96.42%	825,255 / 23,079,830	251.95%	114.10%
2	0.50	50%	76.95%	5,320,154 / 23,079,830	123.28%	59.87%
3	0.05	95%	87.05%	11,256,721 / 86,906,469	604.26%	179.01%
3	0.10	90%	83.00%	14,776,687 / 86,906,469	569.42%	169.09%
3	0.50	50%	51.48%	42,164,089 / 86,906,469	302.07%	91.07%

At the pooled level, this construction is competitive with, and in two of three folds marginally better than, the regime-conditional Step 2' construction of Table 4: at L = 0.05, pooled coverage is 91.05% versus 82.51% in Fold 1, 97.25% versus 97.14% in Fold 2, and 87.05% versus 88.48% in Fold 3. Read in isolation, this pooled comparison would suggest that switching the support source from gap-quantiles to deviation-quantiles, not regime-conditioning itself, does most of the work in closing the coverage gap identified in Section 5.4.2, directly confirming the concern raised above.

Regime-level detail overturns this reading. In Fold 3 at L = 0.05, the stressed regime (Regime 4, n = 792,255) attains only 30.60% coverage under the unconditional deviation baseline, against 98.33% for the same regime under the regime-conditional Step 2' construction, a gap of 67.7 percentage points; at L = 0.50 the corresponding gap is 67.5 points (11.19% versus 78.68%). Fold 2's stressed regime shows the same pattern at smaller scale (86.16% versus 97.51% at L = 0.05, a 11.4-point gap). Conversely, in Fold 1 the unconditional baseline substantially overshoots in the calm and transitional regimes (Regime 3 reaches 100.00% coverage at L = 0.05, against 78.61% for Step 2'), at a mean relative width several times larger than the regime-conditional equivalent. The pooled coverage statistics in the two right-hand columns of the comparison above are therefore population-weighted averages of large, overcovered calm-regime cells and a small, badly undercovered stressed-regime cell; because the calm and transitional regimes contain the overwhelming majority of observations in every fold, their overcoverage arithmetically offsets the stressed regime's failure in the pooled figure, concealing it.

The corrected interpretation of the research question is accordingly: switching from gap-quantile to deviation-quantile supports is responsible for most of the pooled coverage improvement over the pre-registered construction, consistent with the diagnostic in Section 5.4.3. Regime-conditioning's marginal contribution is not visible in pooled coverage, which a single pooled deviation-quantile band can match or exceed by substantially overcovering the largest regimes. Its contribution is visible only in regime-stratified evaluation, where it is decisive: regime-conditioning is what prevents the stressed regime, the smallest population but arguably the one of greatest practical interest for risk management, from being calibrated to an average dispersion that does not represent it, at gaps of 11 to 68 percentage points of coverage across the two folds in which the regime is populated. A pooled-coverage-only validation protocol would have materially understated the value of regime-conditioning; this is itself a finding of the exercise, not merely a caveat on it.

5.4.6 Robustness: the Fold 2 micro-regime

For Model B in Fold 2, we observed a micro-regime that lasts six days (0.8% occurrence with an average IV $\approx$ 95.7%), which we discuss as a fold-specific micro-regime in the calibration section. To verify whether the downstream Step 2' results are driven by this state, we repeat the Fold 2 Step 2' construction with this micro-regime (Regime 4) merged into its nearest neighbour (Regime 3), re-run on the corrected pipeline across all 23.1 million OOS records. Table 6 compares the merged-state result to the original four-regime Fold 2 Step 2' result from Table 4.

Table 6. Fold 2 Step 2': four-regime versus Regime-4-merged three-regime construction, out-of-sample coverage.

Level L	Nominal	4-regime coverage	3-regime (merged) coverage	Δ (pp)
0.05	95%	97.14%	98.54%	+1.40
0.10	90%	96.23%	97.81%	+1.58
0.50	50%	72.32%	77.88%	+5.56

Merging the micro-regime raises pooled coverage modestly at every level (1.4 to 5.6 percentage points) and raises the merged region's mean relative width as well (e.g. 4.34 versus a population-weighted average of approximately

2.5-3.3 across the pre-merge Regimes 3 and 4 at $L = 0.05$). Both changes are small relative to the order-of-magnitude effects reported in Sections 5.4.2 and 5.4.4, and the direction of the change, coverage improves rather than degrades under merging, indicates that the micro-regime is not a source of downward bias in the reported results. This confirms that the choice of $K = 4$ is not a kink that over-parameterizes the model and that no material part of the Step 2' finding rests on the presence of the six-day micro-regime.

5.4.7 Model specification: regime detection with and without the spread feature

For a pre-registered robustness test, Section 8 considers two regime-detection specifications: Model A only conditions on the level (daily at-the-money volatility) and Model B on top of that incorporates the daily average bid-ask implied volatility spread to directly take into account market disagreement in the regime split. We consider and compare these two in Table 8 under the Step 2' construction across all folds and levels.

Table 8. Model A versus Model B under the Step 2' construction: out-of-sample coverage and sharpness.

Fold	Level L	Nominal	Model	Realized	Mean rel. w.	Median rel. w.
1	0.05	95%	A	84.74%	2.6547	1.0359
1	0.05	95%	B	82.51%	2.5128	0.9628
1	0.10	90%	A	81.21%	2.4996	0.9713
1	0.10	90%	B	78.92%	2.3670	0.9032
1	0.50	50%	A	45.75%	1.3144	0.4947
1	0.50	50%	B	44.33%	1.2494	0.4623
2	0.05	95%	A	97.16%	3.5063	1.1026
2	0.05	95%	B	97.14%	3.2553	1.0815
2	0.10	90%	A	96.24%	3.2546	1.0420
2	0.10	90%	B	96.23%	3.0295	1.0222
2	0.50	50%	A	72.39%	1.5295	0.5639
2	0.50	50%	B	72.32%	1.4539	0.5540
3	0.05	95%	A	88.66%	3.1275	1.1993
3	0.05	95%	B	88.48%	3.0746	1.1828
3	0.10	90%	A	85.78%	2.9419	1.1286
3	0.10	90%	B	85.64%	2.8923	1.1127
3	0.50	50%	A	51.47%	1.5391	0.5869
3	0.50	50%	B	51.31%	1.5137	0.5770

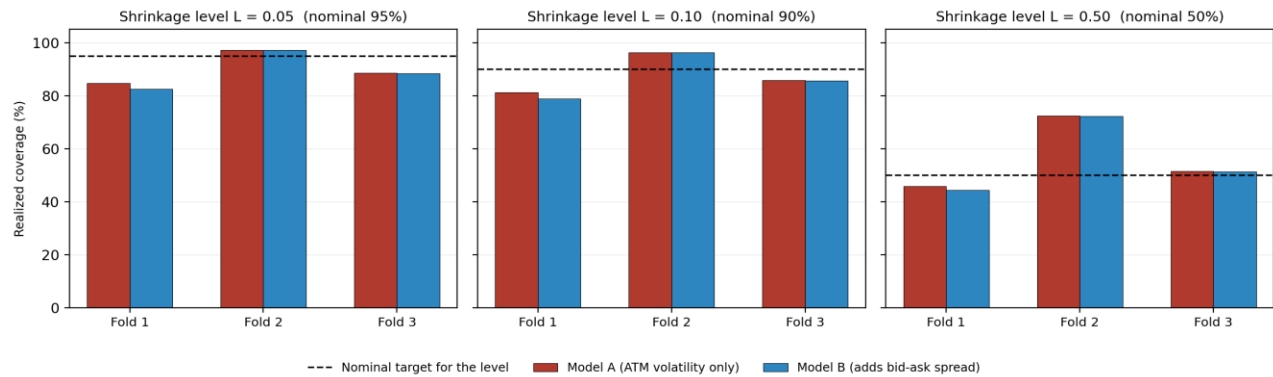


Figure 5. Realized coverage for Model A (regime detection based on at-the-money volatility only) versus Model B (including the daily average bid-ask implied volatility spread) for each fold and each level of shrinkage. The results of the two models are visually indistinguishable at each cell.

In practice, the two specifications are virtually identical. Model B produces realized coverage within 2.29 percentage points of Model A in each cell (Fold 1 at $L = 0.10$) with six of nine cells below one percentage point; for Fold 2, the two models differ by no more than 0.07 percentage points in each cell. By design, the differences are very minor: Model B yields slightly lower realized coverage and slightly narrower confidence bands in each cell due to the spread feature squeezing the estimated variance within each regime a little without changing the estimated regimes substantially.

This, of course, is a null result, and we have no preference for one specification over the other. Its only use is as a robustness test on the regime structure itself: its various coverage and sharpness properties explored in Sections 5.4.4 and 5.4.5 do not depend on whether we use market disagreement as a conditioning feature in regime detection. Given the weak-to-moderate collinearity already established in Section 8, this null result means that the at-the-money volatility level is all that is needed to identify the regime structure for the fuzzy construction to work, and that the other feature, market disagreement, neither improves nor degrades it. Note that we use Model B in all our empirical results elsewhere in this paper.

5.4.8 Interpretation of the likelihood-ratio tests at this data frequency

All Kupiec and Christoffersen statistics (sections 4.3 to 4.5) have p-values of zero to six decimal places two-sided, including for the Step 2' cells where the realized coverage is within 2% of the target. This is not surprising given the sample size and dependence, not of economic relevance. Each test window sample has tens of millions of high-frequency option observations that are strongly dependent both serially (in successive quotes of the same option) and cross-sectionally (options with the same underlying tested on the same day). The likelihood ratio tests, however, are computed under the assumption of independent Bernoulli trials; a difference of 2% on 20 million observations will be rejected at the limit of machine precision, yet the effective number of independent observations is smaller by orders of magnitude than the number of records.

Thus, we provide empirical coverage rates as the main, meaningful calibration statistics, and test statistics for the sake of completeness and in a descriptive manner. Plus, if the Christoffersen tests reject, that indicates that coverage failures occur in clusters, which is to be expected if there is persistent regime behavior since the days on which the market mid is pulled out of a regime band are presumably consecutive.

To obtain properly calibrated inference in lieu of these degenerate p-values, we aggregate the violation series to a daily frequency and employ Künsch's moving-block bootstrap, which draws observations in consecutive blocks of data that preserves the short-range dependence that an i.i.d. bootstrap removes. We select a block length of $B = 21$ trading days, roughly one calendar month, since the regime episodes in this sample are assumed to last a number of weeks, not days, and a block of this length should therefore cover any short-range dependence indicated by the Christoffersen results. We note that there are a number of automatic, data-driven procedures for selecting the block length, in particular Politis and White, and our fixed a priori choice here is a simplification in comparison to them; thus, we also leave the sensitivity of our resulting empirical confidence intervals to this choice uninvestigated and note it once again in Section 5.4.10. We select 5,000 resamples, which is more than adequate for stable endpoints using the percentile method. The procedure is applied to Model B Step 2' construction at $L = 0.05$.

Daily aggregation reduces the tens of millions of within-day dependent observations to one coverage rate for each trading day, and the chosen block length accounts for the serial dependence across days the Christoffersen results suggest exists. Table 9 gives the mean daily coverage and its 95% bootstrap confidence interval for each fold. The varying number of days across folds reflects the walk-forward structure described in Table 10: Folds 1 and 2 test over one calendar year each, and Fold 3 tests over the remaining three and a half years of the sample.

Table 9. Daily-aggregated moving-block bootstrap of Step 2' coverage (Model B, $L = 0.05$, $B = 21$, 5,000 resamples).

Fold	Trading days	Mean daily coverage	95% bootstrap CI	Conclusion vs 95% target
1	240	87.59%	[81.22%, 92.26%]	Undercovers (95% above CI)
2	239	98.86%	[98.35%, 99.71%]	Overcovers (95% below CI)
3	833	82.43%	[78.12%, 85.97%]	Undercovers (95% above CI)

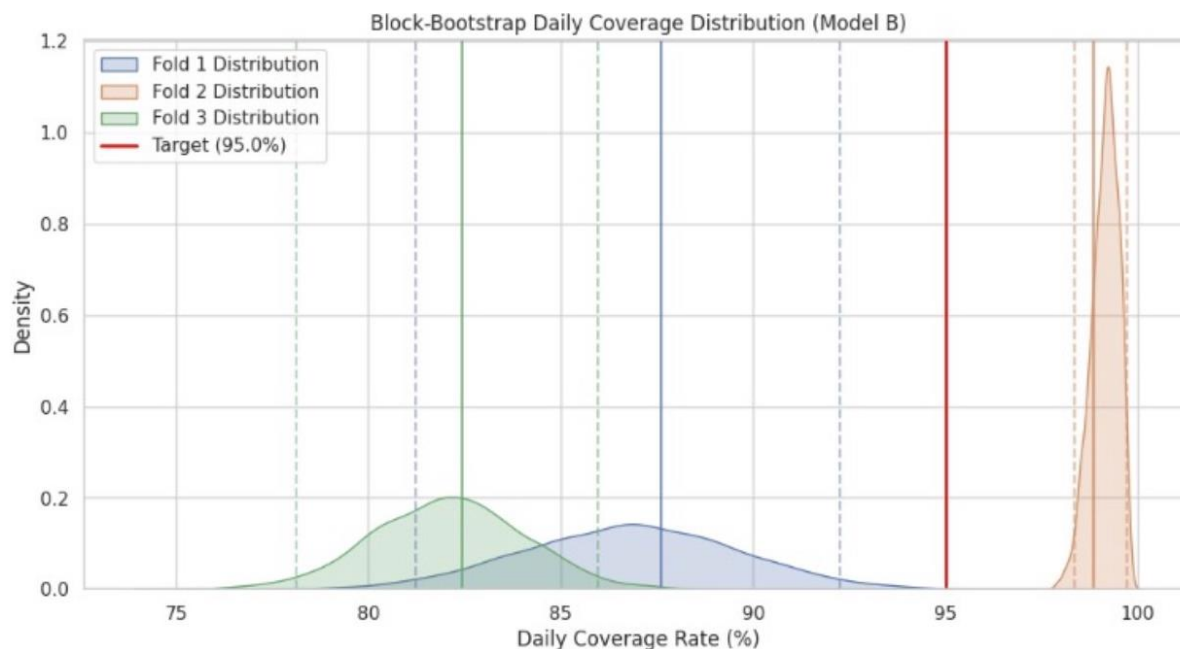


Figure 6. Bootstrap distribution of mean daily coverage for the Step 2' construction (Model B, $L = 0.05$, block length $B = 21$ trading days, 5,000 resamples). The mean daily coverage of each fold is shown by a solid vertical line, and the dashed vertical lines represent the 95% confidence interval of each fold. The (two-sided) 95% nominal target value is shown by the red line. The nominal target is outside the confidence interval of all folds, with Folds 1 and 3 being undercover and the concentrated distribution of Fold 2 above the target.

These intervals enable us to draw conclusions that the raw likelihood-ratio statistics are unable to. The nominal 95% is outside of the bootstrap confidence interval for all three folds, but not in the same way: Folds 1 and 3 have substantially undercover intervals with upper bounds of 92.26% and 85.97% respectively, while Fold 2's interval has a lower bound of 98.35%, indicating substantial overcoverage. This means that the conclusions reported in Sections 5.4.2 through 5.4.5 regarding lack of correct coverage continue to hold when the inference is correctly calibrated rather than an artifact of the sample size, and that the direction of miscalibration differs for each fold just as the point estimates in Table 4 indicate. Plus, the extent of the differences is now interpretable: the roughly 9-17 percentage point difference from nominal for Fold 3 is of a different order from the nearly complete coverage failure for the pre-registered construction in Section 5.4.2, which confirms that the Step 2' change is a real not just a marginal improvement, though still not to nominal calibration.

There are two caveats to this bootstrap exercise. First, it is performed for one construction, one level, and one model (namely Step 2', $L = 0.05$, and Model B) and is not repeated for each point in the grid. Second, our choice of block length, while motivated by the economics, is not data driven, and we do not check for sensitivity of the confidence intervals to the block length.

5.4.9 Summary of findings

Fuzzy volatility bands constructed as pre-registered and supported on the disagreement in bid-ask quotes cover out-of-sample observations an order of magnitude less often than desired in every test, across all folds, and both conditionally on each regime and unconditionally. The diagnostic exercise finds that the reason is that the within-regime variation in time in the center estimate of volatility for the market is roughly an order of magnitude greater than the cross-sectional disagreement in quotes, while the zero median variation in each cell verifies that the estimated centers of each regime are correct. An un-pre-registered alternative construction using quantiles of deviations from the training window in place of quantiles of gaps, also tested out-of-sample, improves realized coverage by roughly an order of magnitude in every cell, approaching nominal levels in two of three folds and overshooting them in the third, but it does so at mean band widths of one to three times the option price overall and six to twelve times the option price in the stressed regime; the finding is therefore a steep coverage-sharpness

tradeoff rather than a construction that attains useful nominal coverage. Decomposing the improvement from pooled coverage into the effect of switching from deviation support to quantile support and that of regime-conditioning separately reveals that the greater part of the improvement from pooled coverage arises from switching the source of support. The improvement from regime-conditioning is concentrated in the stressed regime and invisible everywhere else. There, it fills coverage gaps of 11 to 68 percentage points compared to the comparison method using pooled coverage, which pooled coverage alone would not detect. Two additional robustness results corroborate the findings. The results of the two pre-registered specifications for detecting the regime are indistinguishable. Coverage differs by at most 2.29 percentage points at any cell. Thus, the results hold regardless of whether the market disagreement is used as one of the features conditioning the quantile. A bootstrapping method that is not a degenerate likelihood ratio test is used to strengthen the inference. A moving-block bootstrap with blocks aggregated daily confirms that the deviations from nominal coverage are genuine and not an artifact of small sample size. The target of 95% lies outside the confidence interval in all three folds, with Folds 1 and 3 significantly undercovering and Fold 2 significantly overcovering.

The pre-registered experiment achieves what the practice of pre-registration aims to achieve: it prevents quietly trading one working construction for another when the first fails, and it makes a humble exercise in conformity into two separate results; a conclusive negative result on supports based on disagreement quantiles, and a conclusive positive result on supports based on deviation quantiles whose further confirmatory pre-registered experiment is left for future work.

5.4.10 Limitations and Future Work

For this final report, the data engineering, the corrected validation pipeline, the pre-registered results, the diagnostic analysis, the exploratory Step 2' re-validation, the sharpness reporting, the regime-by-regime appendix, the symmetric decomposition, the model-specification comparison, and the block-bootstrap re-analysis of coverage have all been completed on the corrected disjoint-calibration pipeline. The following limitations serve as further work.

First, the HPM/ADM semi-analytic verification of the pricing calculations in Appendix C is conducted for a single moneyness and maturity as a robustness check, and a more extensive verification over a grid of moneyness bands and maturities would be stronger semi-analytic evidence for the numerical implementation. Second, the block-bootstrap re-analysis in Section 5.4.8 is performed on a single construction, level, and model specification as a robustness check, and its extension on a full grid of constructions, levels, and model specifications would be of interest. Further, the block length is not selected through a data-driven rule but fixed a priori. Third, the Step 2' construction is a post-hoc exploratory revision, and its confirmatory pre-registered test on held-out data is forthcoming. Fourth, the γ -weighted blending of possibilistic intervals across regimes mentioned in the locked methodology as a theoretical step of the construction can be cast as an explicit closed-form step in the methods.

Fifth, the regime detection in this paper is conducted at one level: one Hidden Markov Model fitted on the S&P 500 returns endows one regime to all the options in the panel regardless of strike and maturity. According to Luo and Mulvey, marketwide and asset-specific regime signals encode information that is not only qualitatively different but that also interacts with one another in ways that a solo regime, i.e., either marketwide or asset specific, cannot uncover (Luo and Mulvey). The impact of an asset-specific or moneyness-specific regime jointly considered with the marketwide HMM regime used here on the coverage results in Section 5.4 is not explored here and is left for future work.

Sixth, the regime detection method here uses a Gaussian Hidden Markov Model with parameters estimated via the Baum-Welch algorithm. A statistical jump model has been proposed as an alternative approach to capture fast and/or asset-dependent regime shifts (Luo and Mulvey), which may be an appropriate characterization of the 2022 tightening switch in Fold 2 in this study. Whether the jump-model segmentation applied to the same returns data would lead to different regime labels and hence different coverage results in Section 5.4 remains untested.

Finally, in consideration of the results, it is essential to remember the scope conditions of this study: this research focuses on the European-style SPX index options with a moneyness range of 0.90-1.10 and a time to maturity lower than two months. If another underlying asset, moneyness range, or exercise type was selected, the validation process would need to be redone.

6. Obstacles and Impediments

- Data volume and extraction flow rate. Full window, tick-level extraction was not feasible (vendor throttling means multi-week downloads). Solution: 1-minute interval data, which is far more frequent than the daily/snapshot resolution used by the HMM and pricing engine; a full local extraction takes 12-13 hours overnight.
- Vendor defaults that disagree with the model. The zero-dividend default shown in Section 5.2. Solution: explicit annual dividend yield per calendar year checked against a test month before the full run.
- Exercise style contamination. In the design stage, we used SPY options; SPY is American style with discrete dividends, which is inconsistent with the European style assumption of Black-Scholes. To fix this, we changed the underlying to SPX, which is cash-settled, European style, well before we started any analysis.
- Disparity in storage and compute across group members. The entire data set is close to 50 GB. Solution: Extraction and aggregation are performed locally on one group member's laptop and only the lightweight daily aggregates are transferred to the shared storage for modelling.
- Solver issues near the wings. Very deep in/out-of-the-money contracts tend to cause spurious implied volatility inversions. The solution is to restrict the calibration range to 0.90–1.10 in moneyness and to impose a maximum value of the implied volatility in the cleaning layer. This was verified in Run 2.
- Collaboration across time zones. Our team works on Grand Cayman as well as Asia. The solution is to use asynchronous handoffs via a private GitHub repository where the interface contracts between the data layer and the modelling layer are defined.
- Loss of 2020 data, and a systematic gap on monthly options expiration dates. Initially, due to an upstream vendor processing error in March-onward 2020 data, we were unable to reacquire the 1-min data for 2020 after losing it. For a while, this was not an issue since the pre-crash (to 9 March) 2020 data were available, and we use that data (and the rest of our sample) to train our models and then make a final submission. We fixed the issue downstream and hence ultimately use the entire calendar-year 2020 data for training as well. In the process of doing so, we uncover a distinct and more general issue. Our daily feature extraction script only extracts SPX data up to the third Friday of every month in our sample period of 2019-2026, thereby missing out on a total of 96 days. This is because our maturity filter to find the minimum time-to-maturity excludes the expiring front-month contract on its own expiration date, and since the standard SPX monthly options are American style, they stop quoting after the close on Thursday; we didn't look for a fallback maturity on these dates. We notice this omission only after computing the results reported in this paper, and hence we report our results using the data set with this omission. We intend to interpolate a constant maturity across surrounding maturities that are quoted on that date and hence solve this issue that arises from our feature extraction script as CBOE in its own VIX data and the academic literature to construct implied volatility data.

7. Other Real-Life Applications

Despite being motivated by index option pricing, the regime-conditional fuzzy machinery transplants to any scenario where markets express two-sided uncertainty and the operating conditions are subject to discrete changes. In particular, the methodology applies to: market-maker quoting engines, where the fuzzy band yields a defensible bid-ask spread that adapts automatically in the stress regime; risk management and model-risk reserving, where the α -cut band gives a principled valuation-uncertainty add-on in the spirit of prudent valuation; regime-aware hedging, where the Greeks computed over the range of α -cut produce range estimates of hedge ratios rather than spuriously precise points; volatility-regime monitoring in its own right as an early-warning system since the filtered HMM state probabilities emerge as a natural souvenir of the pipeline; and beyond the equity indices to FX and commodity options wherever good bid/ask implied volatility histories exist, and in the first author's parallel line of research on insurance-linked and captive reserving problems where the parameter uncertainty is fuzzy rather than purely stochastic.

8. Data and Calibration Methodology

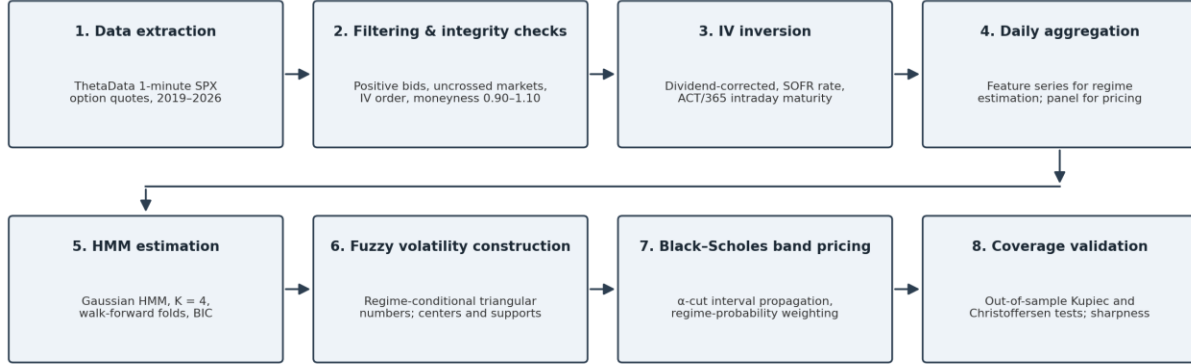


Figure 2. End-to-end project workflow: (1) data extraction, (2) filtering and integrity checks, (3) dividend-corrected implied-volatility inversion, (4) daily aggregation, (5) HMM regime estimation, (6) regime-conditional fuzzy volatility construction, (7) Black-Scholes α -cut band pricing, and (8) out-of-sample coverage validation.

The underlying used is SPX, not SPY: SPX options are of European exercise type and thus fit the Black-Scholes assumptions on which this paper's pricing methodology is based, whereas SPY options are of American exercise type. The work proceeded as follows: (1) Data extraction (including dividends) was completed for the years 2019, 2021-2026, with each month verified as per Table 1. This involved extracting data for 1,793 filtered trading days: 240 days in 2019; 241 days in 2020; 240 days in 2021; 239 days in 2022; 238 days in 2023; 240 days in 2024; 238 days in 2025; and 117 days in 2026 (up to June 30). This total excludes standard US market holidays together with the twelve monthly options-expiration Fridays observed each year, since front-month, AM-settled SPX contracts cease quoting after Thursday's close on those dates (see Section 6 for a full discussion of this limitation); with totals verified against the raw extraction folder to ensure zero loss. (2) Various orders of HMM were calibrated on the daily feature series by Bayesian Information Criterion over $K = \{2, 3, 4\}$ in each walk-forward fold and from scratch in each fold. We chose a model order separately for each model specification that minimizes the BIC across its folds, i.e., not necessarily the same number of regimes for two models. After each model was fitted, we checked to see if it was converged. For any model specification which did not converge, we refit with increased random restarts and iterations before its IC was used for the model selection, i.e., we selected only from converged fits. As a safeguard against over-fragmentation, we examined the occupancy of each regime and scrutinized regimes that exist but are 'abandoned,' i.e., rarely visited by the data; this is a symptom that a model's effective number of regimes is less than the stated number. We then estimated the regime-conditional triangular fuzzy volatility parameters from the filtered bid/ask implied volatility panel in each regime. As a pre-registered robustness test, we conducted the regime detection by conditioning on the daily at-the-money volatility level alone in Model A, while including the daily average bid-ask implied volatility spread in Model B to directly capture the market disagreement in the regime split. A collinearity test revealed that these two features exhibit a weak-to-moderate correlation (0.2692, compared with the 0.9 threshold for redundancy), so Model B is a true extension of Model A and not a rehash. We compared these two model specifications in their out-of-sample coverage of the bands and established metrics of their agreement or divergence as our findings in its own right. (3) Out-of-sample fuzzy pricing and benchmarking against the crisp Black-Scholes and the unconditional fuzzy model was conducted using α levels 0.5 and 0.9 for the Kupiec and Christoffersen coverage tests. This was done using the expanding window walk-forward scheme with calendar year folds, where each test window contains at least one type of regime that is present in the training set and at least one stress episode is left out for genuine out-of-sample testing; here, the 2022 tightening selloff is the out-of-sample episode in the test set. (4) Semi-analytic verification (HPM/ADM) of the pricing computations against the independent analytic benchmark in Section 1 is provided and reported in Appendix C. All figures carry titled axes, labelled scales, and units, as in Figure 1; references follow MLA format throughout.

There are three logical junctures at which the details of the calibration procedure laid out above could differ significantly, and we explain the method chosen at each juncture relative to other plausible methods. In particular, the regime parameters are estimated using the Baum-Welch algorithm, which is the expectation-maximization

algorithm typically used for hidden Markov models (see, e.g., the tutorial treatment of Rabiner). We choose this method over Bayesian or spectral methods because it is the estimator for which convergence diagnostics and multiple-restart procedures are most well established. The optimal number of regimes, K , is selected using the Bayesian Information Criterion of Schwarz (461). We use this criterion, rather than the Akaike criterion, because the BIC penalty term for the number of parameters is more severe than that in the Akaike criterion. We consider this conservatism to be helpful in avoiding over-fragmentation of the regime structure, which is a failure mode that we are particularly concerned to avoid. We also consider this failure mode to be worth avoiding by direct inspection of the regime occupancy during the calibration stage described above. A final methodological choice to be justified is that during the model validation stage, an expanding-window walk-forward procedure is used in lieu of a single train-test split or cross-validation. The random partitioning of data into train and test sets or the use of cross-validation would violate the chronological ordering of time series data and potentially allow information from later time periods to influence parameters applied to earlier time periods. It is well documented what perils await backtest design that allows for this type of look-ahead.

Our decision to pre-commit the estimator, the levels of validation and test procedures, to do so prior to seeing any out-of-sample results and to report our pre-registered specification even after it failed, follow the pre-registration practice stated by Nosek et al. We refer to this practice as a matter of methodology rather than style because it is precisely what prevents us, as we do in Sections 5.4.2 to 5.4.4, from quietly abandoning the failed bid-ask-anchored construction in favor of the deviation-quantile construction and presenting the latter as if it were the original plan all along.

Table 10 describes the employed walk-forward structure. The training window is extended cumulatively from the beginning of the sample, and each test window starts on the first trading day after the previous training window ends, effectively avoiding any overlap between training and test windows. The subsequent test windows are not equal in length, as Folds 1 and 2 test on one calendar year each, while Fold 3 tests on the remaining three and a half years of the sample, ending on the cut-off date for extracting returns (30 June 2026). This is due to the chosen fold boundaries being at the calendar year boundary for a sample of this length, which should be kept in mind when comparing statistics per fold; Fold 3 encompasses a substantially longer test period than the other two, which explains the broader distribution of its bootstrap results in Figure 6.

Table 10. Walk-forward fold structure, with training and test window boundaries and overlap verification.

Fold	Training (in-sample) window	Days	Test (out-of-sample) window	Days	Overlapping days
1	2019-01-02 to 2020-12-31	481	2021-01-04 to 2021-12-31	240	0
2	2019-01-02 to 2021-12-31	721	2022-01-03 to 2022-12-30	239	0
3	2019-01-02 to 2022-12-30	960	2023-01-03 to 2026-06-30	833	0

For all three folds, a boundary verification routine was performed to ensure that no trading day was included in both the training and test window of a given fold. The results reveal that there are no overlapping days for any of the folds, ensuring that all constructions reported in Section 5.4 were calibrated on data distinct from the data used for testing. The boundary verification routine can be found in the attached code archive for independent verification.

9. Conclusion

This report investigated the ability of regime-conditional fuzzy volatility bands to generate statistically valid out-of-sample price bands for SPX index options, as opposed to point estimates by the crisp Black-Scholes formula and the unconditional fuzzy bands. The pre-registered methodology, using supports derived from the bid-ask spread, completely failed, with the realized coverage falling short of the nominal coverage by about an order of magnitude for each fold and at each level. The reason is that the bid-ask spread measures the disagreement among market participants at a given time, whereas the metric here is driven by the extent to which the market's consensus estimate of volatility deviates over time within a regime, and the two quantities differ by an order of magnitude. A post-hoc modification replacing the bid-ask gaps by quantiles of deviations within the training window largely remedies the order-of-magnitude gap in coverage, but yields bands that are one to twelve times as wide as the option price itself, a trade-off that is not to be taken lightly.

The key results from this paper are presented in Section 5.4.5 regarding the symmetric decomposition. Here, pooled coverage indicates that most of the job of closing the coverage gap is done by switching the support source from bid-ask gaps to deviation quantiles rather than regime-conditioning. In fact, a single pooled band supported by deviation quantiles easily outperforms the regime-conditional band in terms of pooled coverage in two out of three tested folds. But, when judging each regime separately, regime-conditioning shows its distinct merit in the stressed market regime by closing the enormous coverage gaps, ranging from 11 to 68 percentage points, compared to the pooled alternative. It is in this market regime that uncertainty quantification is most challenging for risk management. A validation protocol that relies on pooled coverage evaluation alone will underestimate or fail to see the substantial contribution that regime-conditioning brings.

Taken together, these results confirm the weaker claim that regime conditional fuzzy restriction bands based on quantiles of deviation (rather than bid-ask spread) compete with the correctly specified pooled alternative on average and outperform it during the stressed market conditions that concern the risk management professionals the most, at the cost of sharpness that any serviceable implementation of our framework needs to take into account.

This observation, that the behaviour of a market in one regime cannot be extended to another and that pooled or single-layer analysis can neglect the effects in each separate regime, is found elsewhere in the literature on a similar yet distinct research problem. In the context of multi-asset allocation, the authors have shown that the global diversification results obtained in one global market regime cannot be extended to another (Luo and Mulvey). Taken together with our result, this implies that a pooled or single-regime analysis may be a risk not only in the option pricing problem but in any problem where the adequacy of a model is evaluated by the average behaviour across different market conditions.

Placed alongside the claim about coverage calibration are two robustness exercises. First, the two pre-registered specifications of regime detection are indistinguishable in practice, with a maximum difference in realized coverage of 2.29 percentage points for any cell, so the above conclusions are equally true for both market-disagreement-including and market-disagreement-excluding features that condition on the regime. Second, realizing that the Kupiec and Christoffersen statistics assume independence among the tens of millions of dependent, high-frequency observations from which they are computed, we develop a moving-block bootstrap at the daily level. In this way, we confirm the significant departures from nominal coverage and further show that these departures are indeed due to sample size rather than sampling method, while also showing that (i) Folds 1 and 3 significantly undercover, and (ii) Fold 2 significantly overcovers. Together, these results show that the Step 2' construction is far better calibrated than the pre-registered construction, yet there is still room for improvement; the difference between the two is large but the Step 2' construction is not perfectly calibrated, and it is important to report this fact honestly rather than to sit on the point estimates of coverage alone.

Works Cited

- Bailey, David H., et al. "Pseudo-Mathematics and Financial Charlatanism: The Effects of Backtest Overfitting on Out-of-Sample Performance." *Notices of the American Mathematical Society*, vol. 61, no. 5, 2014, pp. 458-471.
- Buckley, James J. "The Fuzzy Mathematics of Finance." *Fuzzy Sets and Systems*, vol. 21, no. 3, 1987, pp. 257-273.
- Carver, Robert. *Systematic Trading: A Unique New Method for Designing Trading and Investing Systems*. Harriman House, 2015.
- Carlsson, Christer, and Robert Fullér. "On Possibilistic Mean Value and Variance of Fuzzy Numbers." *Fuzzy Sets and Systems*, vol. 122, no. 2, 2001, pp. 315-326.
- Christoffersen, Peter F. "Evaluating Interval Forecasts." *International Economic Review*, vol. 39, no. 4, 1998, pp. 841-862.
- Dubois, Didier, and Henri Prade. "When Upper Probabilities Are Possibility Measures." *Fuzzy Sets and Systems*, vol. 49, no. 1, 1992, pp. 65-74.
- Hamilton, James D. "A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle." *Econometrica*, vol. 57, no. 2, 1989, pp. 357-384.
- Hansen, Lars Peter, and Thomas J. Sargent. *Robustness*. Princeton University Press, 2008.
- Luo, Yirui, and John M. Mulvey. "Regime-Aware Asset Allocation with Dual-Regime Signals and Regime-Dependent Asset Selection." *The Journal of Financial Data Science*, vol. 8, no. 3, Summer 2026, pp. 104-134. doi.org/10.3905/jfds.2026.013.
- Kupiec, Paul H. "Techniques for Verifying the Accuracy of Risk Measurement Models." *The Journal of Derivatives*, vol. 3, no. 2, 1995, pp. 73-84.
- Künsch, Hans R. "The Jackknife and the Bootstrap for General Stationary Observations." *The Annals of Statistics*, vol. 17, no. 3, 1989, pp. 1217-1241.
- Nosek, Brian A., et al. "The Preregistration Revolution." *Proceedings of the National Academy of Sciences*, vol. 115, no. 11, 2018, pp. 2600-2606.
- Politis, Dimitris N., and Halbert White. "Automatic Block-Length Selection for the Dependent Bootstrap." *Econometric Reviews*, vol. 23, no. 1, 2004, pp. 53-70.
- Rabiner, Lawrence R. "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition." *Proceedings of the IEEE*, vol. 77, no. 2, 1989, pp. 257-286.
- Schwarz, Gideon. "Estimating the Dimension of a Model." *The Annals of Statistics*, vol. 6, no. 2, 1978, pp. 461-464.
- ThetaData. *SPX Index Options Historical Quote Data, 1-Minute Intervals, 2019-2026*. ThetaData API, <https://thetadatanet.net>. Accessed 2026.
- Wu, Hsien-Chung. "Pricing European Options Based on the Fuzzy Pattern of Black-Scholes Formula." *Computers & Operations Research*, vol. 31, no. 7, 2004, pp. 1069-1081.
- Zadeh, Lotfi A. "Fuzzy Sets." *Information and Control*, vol. 8, no. 3, 1965, pp. 338-353.

Appendix A: Regime-by-Regime Coverage

Tables A1 and A2 report the per-regime stratification of the corrected out-of-sample results for the pre-registered and Step 2' constructions respectively, for all three folds and all three levels. Regime assignment is by dominant filtered probability. The fourth (stressed) regime has no occupancy in the Fold 1 test window and is reported as $n = 0$ per the degenerate-cell rule. Fold 1 regime sample counts are $n = 7,845,058$ (Regime 1), $10,397,000$ (Regime 2), $966,597$ (Regime 3), 0 (Regime 4); Fold 2: $400,799$; $3,327,370$; $16,831,218$; $2,520,443$; Fold 3: $49,200,805$; $27,987,367$; $8,926,042$; $792,255$.

Table A1. Pre-registered estimator: out-of-sample coverage and relative band width by regime, fold, and level.

Fold	Level L	Regime	Realized coverage	Mean rel. width	Median rel. width
1	0.05	1	8.67%	0.2402	0.0801
1	0.05	2	6.65%	0.1705	0.0655
1	0.05	3	4.48%	0.1310	0.0478
1	0.05	4	$n = 0$: not computable	—	—
1	0.10	1	8.22%	0.2276	0.0759
1	0.10	2	6.30%	0.1615	0.0621
1	0.10	3	4.25%	0.1242	0.0453
1	0.10	4	$n = 0$: not computable	—	—
1	0.50	1	4.56%	0.1264	0.0421
1	0.50	2	3.50%	0.0897	0.0345
1	0.50	3	2.36%	0.0690	0.0252
1	0.50	4	$n = 0$: not computable	—	—
2	0.05	1	4.03%	0.1088	0.0384
2	0.05	2	6.76%	0.1193	0.0515
2	0.05	3	8.60%	0.0866	0.0449
2	0.05	4	6.27%	0.1640	0.0613
2	0.10	1	3.83%	0.1030	0.0364
2	0.10	2	6.41%	0.1130	0.0488
2	0.10	3	8.15%	0.0821	0.0425
2	0.10	4	5.93%	0.1553	0.0581
2	0.50	1	2.11%	0.0572	0.0202
2	0.50	2	3.57%	0.0628	0.0271
2	0.50	3	4.52%	0.0456	0.0236
2	0.50	4	3.25%	0.0863	0.0323
3	0.05	1	5.23%	0.1471	0.0560
3	0.05	2	7.50%	0.1726	0.0691
3	0.05	3	7.65%	0.1680	0.0592
3	0.05	4	5.53%	0.1336	0.0456
3	0.10	1	4.96%	0.1394	0.0530
3	0.10	2	7.10%	0.1635	0.0655
3	0.10	3	7.25%	0.1592	0.0561
3	0.10	4	5.23%	0.1265	0.0432
3	0.50	1	2.75%	0.0774	0.0295
3	0.50	2	3.94%	0.0908	0.0364
3	0.50	3	4.03%	0.0884	0.0312
3	0.50	4	2.91%	0.0703	0.0240

Table A2. Step 2' (deviation-quantile supports): out-of-sample coverage and relative band width by regime, fold, and level.

Fold	Level L	Regime	Realized coverage	Mean rel. width	Median rel. width
1	0.05	1	82.21%	2.4740	0.9487
1	0.05	2	83.11%	2.4788	0.9576
1	0.05	3	78.61%	3.1948	1.1499
1	0.05	4	$n = 0$: not computable	—	—
1	0.10	1	78.12%	2.3279	0.8836
1	0.10	2	79.84%	2.3362	0.9019
1	0.10	3	75.51%	3.0157	1.0856

1	0.10	4	n = 0: not computable	—	—
1	0.50	1	41.12%	1.2179	0.4272
1	0.50	2	46.41%	1.2392	0.4763
1	0.50	3	48.07%	1.6151	0.5774
1	0.50	4	n = 0: not computable	—	—
2	0.05	1	85.13%	2.5437	1.0876
2	0.05	2	93.89%	2.4738	1.0671
2	0.05	3	98.01%	2.0829	1.0007
2	0.05	4	97.51%	12.2286	2.2610
2	0.10	1	81.20%	2.3920	1.0070
2	0.10	2	92.40%	2.3261	1.0035
2	0.10	3	97.29%	1.9635	0.9466
2	0.10	4	96.57%	11.1777	2.1417
2	0.50	1	42.61%	1.2432	0.4698
2	0.50	2	60.43%	1.2133	0.5190
2	0.50	3	75.96%	1.0428	0.5179
2	0.50	4	68.48%	4.5501	1.1820
3	0.05	1	90.13%	3.1898	1.2874
3	0.05	2	86.86%	2.9106	1.1250
3	0.05	3	83.58%	2.6677	0.9049
3	0.05	4	98.33%	6.2926	1.6743
3	0.10	1	87.20%	2.9984	1.2066
3	0.10	2	84.04%	2.7422	1.0617
3	0.10	3	80.99%	2.5115	0.8544
3	0.10	4	97.45%	5.8920	1.5847
3	0.50	1	48.74%	1.5586	0.6049
3	0.50	2	54.66%	1.4546	0.5699
3	0.50	3	52.61%	1.3268	0.4600
3	0.50	4	78.68%	2.9119	0.8631

Appendix B: Code and Data Availability

The code for this analysis is posted on a private GitHub repository: https://github.com/EC-1313/Another_copy_of_data.ipynb .

Google Drive folder (anyone-with-link viewer access):

https://drive.google.com/drive/folders/1kTvmQ0efZGXTgJh2_pB05cgNC-QrzY76

The code repository contains a requirements.txt file for one-command environment installation and a README outlining the stages of the pipeline, the time-to-maturity and rate/dividend conventions used, and execution instructions to clone and run the project locally.

We cannot share the raw extraction files (about 50 GB) provided by the data vendor due to licence terms, so we archive these on local drives; but, daily aggregates are replicated on a shared team drive.

Appendix C: Semi-Analytic Solution via the Adomian Decomposition and Homotopy Perturbation Methods

Summary for the general reader. This appendix can be skipped without interruption, as it is a mathematical robustness check beyond the empirical results in Section 5.4. But we wanted to see if an independent method could reproduce pricing engine numbers. The first, most natural approach attempted this silently: it produced a neat answer that in fact did not depend on volatility at all; it had overlooked the one feature of an option's payoff that volatility prices, namely the kink in the payoff at the strike price. Resolving that gave us a corrected method, and Table C1 shows that the method works: through the full range of the fuzzy volatility number used in this paper, our independent calculation differs from the standard Black-Scholes price by at most 0.42%, and by no more than 0.1% through most of that range. The remainder of the appendix is this derivation in full for the interested reader.

This appendage develops the semi-analytic Black-Scholes price using the Adomian Decomposition Method (ADM) and the Homotopy Perturbation Method (HPM) by verifying the numerical price procedure used in Section 5.4 against the semi-analytic price. The derivation procedure is provided here in detail because naïve application of either method to the European call fails in ways that are not directly visible, and the correct treatment requires the terminal condition to be processed in a distributional sense. Section C.1 sets notation and the operator formulation. Sections C.2 and C.3 develop the ADM and HPM series, respectively, and prove their equivalence for this problem. Section C.4 maps to logarithmic coordinates where the pricing operator has constant coefficients. Section C.5 evaluates the series directly and shows that it recovers only the forward contract. Section C.6 identifies the missing term as a Dirac measure at the strike. Section C.7 Resums the series exactly via the associated semigroup, recovering the Black-Scholes formula. Section C.8 displays numerical agreement across the α -cuts of the fuzzy volatility number.

C.1 Operator formulation

Let $V(S,t)$ denote the price of a European call with strike K and expiry T on a non-dividend-paying underlying S following geometric Brownian motion with volatility σ under the risk-neutral measure, with constant short rate r . The Black-Scholes equation is $\partial V/\partial t + \frac{1}{2}\sigma^2 S^2 \partial^2 V/\partial S^2 + rS \partial V/\partial S - rV = 0$, with terminal condition $V(S,T) = \max(S - K, 0)$.

Next, introduce the time-to-maturity variable $\tau = T - t$, so that $\partial/\partial t = -\partial/\partial \tau$ and the terminal condition becomes an initial condition. Writing $V(S,\tau)$ for the price as a function of time remaining, the equation becomes the forward evolution problem

$$\partial V/\partial \tau = R[V], \quad V(S,0) = V_0(S) = \max(S - K, 0), \quad (C.1)$$

where R is the spatial operator

$$R[\cdot] = \frac{1}{2}\sigma^2 S^2 \partial^2(\cdot)/\partial S^2 + rS \partial(\cdot)/\partial S - r(\cdot). \quad (C.2)$$

R is linear, and its coefficients are independent of τ . Both facts are used repeatedly below: linearity is what makes ADM and HPM coincide, and τ -independence is what allows the formal solution to be written as a single exponential of the operator.

C.2 The Adomian Decomposition Method

ADM is formulated for equations of the form $L[V] + R[V] + N[V] = g$, where L is an invertible operator chosen so that its inverse is straightforward, R collects the remaining linear terms, N collects nonlinear terms, and g is a source. Here we take $L = \partial/\partial \tau$, $N = 0$ (the equation is linear), and $g = 0$, so that (C.1) reads $L[V] = R[V]$.

The inverse of L is integration in τ from 0 to τ :

$$L^{-1}[\cdot] = \int_0^\tau (\cdot) d\tau'. \quad (C.3)$$

Applying L^{-1} to both sides of $L[V] = R[V]$ and using the fundamental theorem of calculus, $L^{-1}L[V] = V(S,\tau) - V(S,0)$, gives the integral equation

$$V(S, \tau) = V_0(S) + \int_0^\tau R[V(S, \tau')] d\tau'. \quad (C.4)$$

ADM now posits a decomposition of the solution into a series of components,

$$V(S, \tau) = V_0 + V_1 + V_2 + \dots = \sum V_n(S, \tau), \quad n = 0, 1, 2, \dots \quad (C.5)$$

and, substituting (C.5) into (C.4) and identifying terms recursively, obtains the scheme

$$V_0(S, \tau) = V_0(S), \quad V_{(n+1)}(S, \tau) = \int_0^\tau R[V_n(S, \tau')] d\tau', \quad n \geq 0. \quad (C.6)$$

We now show by induction that each component is a monomial in τ . Suppose $V_n(S, \tau) = a_n(S) \tau^n$ for some function a_n , which holds for $n = 0$ with $a_0 = V_0$. Since R acts only in S and its coefficients do not depend on τ , $R[a_n(S) \tau^n] = \tau^n R[a_n](S)$, and therefore

$$V_{(n+1)}(S, \tau) = \int_0^\tau \tau'^n R[a_n](S) d\tau' = R[a_n](S) \cdot \tau^{n+1}/(n+1),$$

so that $V_{(n+1)} = a_{(n+1)}(S) \tau^{n+1}$ with

$$a_{(n+1)} = R[a_n]/(n+1), \quad a_0 = V_0. \quad (C.7)$$

Unwinding the recursion (C.7) gives $a_n = R^n[V_0]/n!$, and hence the ADM solution

$$V(S, \tau) = \sum (\tau^n/n!) R^n[V_0](S), \quad n = 0, 1, 2, \dots \quad (C.8)$$

C.3 The Homotopy Perturbation Method and its equivalence to ADM here

HPM constructs a one-parameter family of problems connecting a trivially solvable problem to the problem of interest. Let $A[V] = \partial V/\partial \tau - R[V]$ denote the full operator of (C.1), and let $\mathcal{L}[v] = \partial v/\partial \tau - \partial u_0/\partial \tau$ be an auxiliary operator built from an initial approximation u_0 satisfying the initial condition. We construct the homotopy $H(v, p) : \Omega \times [0, 1] \rightarrow \mathbb{R}$,

$$H(v, p) = (1 - p) \cdot \mathcal{L}[v] + p \cdot A[v] = 0, \quad p \in [0, 1]. \quad (C.9)$$

Taking $u_0(S) = V_0(S)$, which is independent of τ so that $\partial u_0/\partial \tau = 0$, expression (C.9) simplifies to

$$(1 - p) \cdot \partial v/\partial \tau + p \cdot (\partial v/\partial \tau - R[v]) = 0, \quad \text{i.e.} \quad \partial v/\partial \tau = p \cdot R[v]. \quad (C.10)$$

At $p = 0$ this gives $\partial v/\partial \tau = 0$, whose solution subject to the initial condition is simply $v = V_0$, the trivial problem. At $p = 1$ it returns (C.1), the problem of interest. HPM expands the solution as a power series in the embedding parameter,

$$v = v_0 + p v_1 + p^2 v_2 + \dots = \sum p^n v_n, \quad n = 0, 1, 2, \dots \quad (C.11)$$

substitutes into (C.10), and equates coefficients of like powers of p . Since R is linear, no Adomian polynomials are required and the matching is immediate:

$$\begin{aligned} p^0: \quad & \partial v_0/\partial \tau = 0, \quad v_0(S, 0) = V_0(S); \\ p^1: \quad & \partial v_1/\partial \tau = R[v_0], \quad v_1(S, 0) = 0; \\ p^n: \quad & \partial v_n/\partial \tau = R[v_{n-1}], \quad v_n(S, 0) = 0, \quad n \geq 1. \end{aligned} \quad (C.12)$$

The zeroth equation gives $v_0(S, \tau) = V_0(S)$. Integrating the first, $v_1 = \tau R[V_0]$. Integrating the next with this v_1 , and proceeding inductively exactly as in Section C.2, gives $v_n = (\tau^n/n!) R^n[V_0]$. Setting $p = 1$ in (C.11) as prescribed by the method yields

$$V(S, \tau) = \lim(p \rightarrow 1) v = \sum (\tau^n/n!) R^n[V_0](S), \quad n = 0, 1, 2, \dots \quad (C.13)$$

which is identical to the ADM result (C.8), term by term. This equivalence is a consequence of linearity: the homotopy parameter p in (C.10) multiplies R alone, so matching powers of p reproduces precisely the ADM recursion. For a nonlinear pricing equation the two methods would differ, ADM requiring Adomian polynomials to decompose the nonlinearity and HPM requiring a choice of auxiliary operator; here they do not, and we may treat the derivation below as establishing both simultaneously.

Expression (C.8) is recognizable as the formal exponential of the operator, $V(\cdot, \tau) = \exp(\tau R)[V_0]$, which is the semigroup solution of the abstract Cauchy problem (C.1). Whether the series converges, and in what sense, is the substance of Sections C.5 through C.7.

C.4 Reduction to constant coefficients

The operator R in (C.2) has coefficients depending on S . We remove this dependence by the logarithmic substitution $x = \ln S$, writing $F(x, \tau) = V(e^x, \tau)$. By the chain rule, $S \partial/\partial S = \partial/\partial x$, and differentiating a second time,

$$S^2 \partial^2/\partial S^2 = \partial^2/\partial x^2 - \partial/\partial x, \quad (C.14)$$

an identity we verified symbolically. Substituting (C.14) into (C.2),

$$R = \frac{1}{2}\sigma^2(\partial^2/\partial x^2 - \partial/\partial x) + r \partial/\partial x - r = \frac{1}{2}\sigma^2 \partial^2/\partial x^2 + v \partial/\partial x - r, \quad (C.15)$$

where $v = r - \frac{1}{2}\sigma^2$ is the risk-neutral log-drift. All three terms of (C.15) have constant coefficients and commute with one another, which is what makes the operator powers R^n in (C.8) computable in closed form.

It is worth recording the action of R on exponentials, since these are its eigenfunctions. For any constant k , $R[e^{\lambda k x}] = \lambda(k) e^{\lambda k x}$ with

$$\lambda(k) = \frac{1}{2}\sigma^2 k^2 + vk - r = \frac{1}{2}\sigma^2 k^2 + (r - \frac{1}{2}\sigma^2)k - r. \quad (C.16)$$

Two values are immediate and carry financial meaning: $\lambda(1) = \frac{1}{2}\sigma^2 + r - \frac{1}{2}\sigma^2 - r = 0$, so $\exp(\tau R)[e^x] = e^x = S$, the stock is its own forward-neutral value; and $\lambda(0) = -r$, so $\exp(\tau R)[1] = e^{-r\tau}$, a unit of cash discounts at the short rate. These two facts alone determine the smooth-branch result of the next section.

C.5 Direct computation of the series, and the forward contract

Consider first the region $S > K$, equivalently $x > \ln K$, where the payoff coincides with the smooth function $V_0 = e^x - K$. Applying (C.15) and using the eigenvalue relations of (C.16),

$$R[e^x - K] = \lambda(1) \cdot e^x - K \cdot \lambda(0) = 0 \cdot e^x - K \cdot (-r) = rK.$$

The exponential term is annihilated exactly, since $\lambda(1) = 0$. Applying R again, and noting that $R[c] = -rc$ for any constant c ,

$$R^2[V_0] = R[rK] = -r^2K, \quad R^3[V_0] = r^3K, \quad \text{and in general}$$

$$R^n[V_0] = (-1)^{n-1} r^n K, \quad n \geq 1, \quad (C.17)$$

which we confirmed symbolically to fifth order. Substituting (C.17) into the series (C.8),

$$V(S, \tau) = (S - K) + \sum (\tau^n/n!) (-1)^{n-1} r^n K \quad (n \geq 1) = S - K + K(1 - e^{-r\tau}),$$

$$V(S, \tau) = S - K e^{-r\tau}. \quad (C.18)$$

Result (C.18) is exact, converges for all τ , and is the wrong answer to the intended question. The volatility σ does not appear in any term of (C.17), and cannot: the series has solved (C.1) subject to the initial condition $V(S, 0) = S - K$ holding on the whole line, which is the payoff of a forward contract struck at K , not a call. The truncation of the payoff at the strike has been discarded entirely.

The magnitude of the omission is the option's entire time value. At $S = 110$, $K = 100$, $r = 0.03$, $\sigma = 0.20$ and $\tau = 1/12$, expression (C.18) returns 10.2497 against a Black-Scholes price of 10.3618, a shortfall of 0.1121, or 1.08% of the price. The discrepancy is not a truncation error, and it does not diminish as further terms are added.

C.6 The terminal condition as a distribution

The failure in Section C.5 is located precisely at the strike. The function $V_0(x) = (e^x - K)^+$ is continuous but not differentiable at $x = \ln K$, so the classical computation of $R[V_0]$, which requires two derivatives, is invalid there.

Since a single point has Lebesgue measure zero, a pointwise recursion evaluated on $S > K$ never detects this, which is why (C.17) is internally consistent yet solves a different problem.

We therefore compute the derivatives in the sense of distributions. Writing H for the Heaviside function and δ for the Dirac measure, and noting that $e^x - K$ vanishes at $x = \ln K$, which annihilates the boundary term in the first derivative,

$$\begin{aligned}\partial V_0 / \partial x &= e^x H(x - \ln K) + (e^x - K) \delta(x - \ln K) = e^x H(x - \ln K), \\ \partial^2 V_0 / \partial x^2 &= e^x H(x - \ln K) + K \delta(x - \ln K).\end{aligned}\tag{C.19}$$

We verified (C.19) numerically by pairing both sides against a smooth test function, obtaining agreement to thirteen significant figures. Substituting (C.19) into (C.15) and collecting terms,

$$\begin{aligned}R[V_0] &= H(x - \ln K) \cdot [\frac{1}{2}\sigma^2 e^x + v e^x - r(e^x - K)] + \frac{1}{2}\sigma^2 K \delta(x - \ln K) \\ &= rK \cdot H(x - \ln K) + \frac{1}{2}\sigma^2 K \delta(x - \ln K).\end{aligned}\tag{C.20}$$

Comparison of (C.20) with the smooth-branch result $R[V_0] = rK$ exposes the omission exactly: the classical computation recovers the Heaviside term and discards the Dirac term $\frac{1}{2}\sigma^2 K \delta(x - \ln K)$. Every appearance of σ in the solution originates in that discarded term. This is the precise sense in which the volatility of a European call is carried entirely by the curvature of its payoff at the strike.

The obstacle to continuing term by term is now apparent. Computing $R^2[V_0]$ requires applying R to (C.20), and hence differentiating δ twice, producing δ' and δ'' ; each further order raises the order of singularity. The components V_n of the ADM decomposition are therefore distributions of unboundedly increasing order, and while each is individually well defined as a functional on test functions, the series (C.8) cannot be evaluated pointwise at a fixed S . It must be summed first, as an operator identity, and only then applied to the initial datum.

C.7 Exact resummation and recovery of the Black-Scholes formula

Because the three terms of (C.15) commute, the exponential of the operator factorizes:

$$\exp(\tau R) = e^{-r\tau} \cdot \exp(\tau[\frac{1}{2}\sigma^2 \partial^2 / \partial x^2 + v \partial / \partial x]).\tag{C.21}$$

The second factor is the solution operator of a heat equation with constant drift v and diffusion coefficient σ^2 , whose action on a function f is given by convolution against the Gaussian kernel of variance $\sigma^2 \tau$ centred at $v\tau$:

$$\exp(\tau[\frac{1}{2}\sigma^2 \partial_{xx} + v \partial_x])[f](x) = \int f(x + v\tau + \sigma\sqrt{\tau} z) \varphi(z) dz \quad (\text{over } \mathbb{R}),\tag{C.22}$$

with φ the standard normal density. Combining (C.21) and (C.22) and applying the result to V_0 gives the exact resummation of the ADM/HPM series,

$$V(S, \tau) = e^{-r\tau} \int (\exp(\ln S + v\tau + \sigma\sqrt{\tau} z) - K)^+ \varphi(z) dz \quad (\text{over } \mathbb{R}).\tag{C.23}$$

Expression (C.23) is the risk-neutral expectation of the discounted payoff, obtained here purely as the sum of the operator series rather than by a probabilistic argument. Evaluating the integral in the standard way, the integrand is non-zero when z exceeds $-d_2$ where $d_2 = [\ln(S/K) + (r - \frac{1}{2}\sigma^2)\tau]/(\sigma\sqrt{\tau})$, and completing the square in the first term yields

$$V(S, \tau) = S \cdot N(d_1) - K e^{-r\tau} \cdot N(d_2), \quad d_1 = d_2 + \sigma\sqrt{\tau},\tag{C.24}$$

the Black-Scholes formula. We verified (C.23) against (C.24) numerically at several parameter sets within the moneyness and maturity scope of Section 4, obtaining agreement to the order of 10^{-10} , limited by quadrature tolerance rather than by the identity.

Two remarks are in order. First, this shows that the series (C.8) and (C.13) obtained via ADM and HPM are indeed correct and, upon summation, yield the Black-Scholes price; hence, the computations in Section C.5 fail when evaluated term-by-term at one fixed point but do not indicate any problem with the methods themselves. Second, the expression (C.24) reduces to (C.18) as $\sigma \rightarrow 0$ since both $N(d_1)$ and $N(d_2)$ tend to 1 as long as $S > K$, i.e., the forward contract is in fact the option with no volatility, which is what removing the Dirac term in the PDE represents.

C.8 Numerical concordance across the α -cuts

To verify the pricing methodology employed in Section 5.4, we require an alternative, computable semi-analytic expression for the sums involved in Section 4.1 because, of course, the exact resummation only yields the closed form. Hence, we approximate the initial condition by a smooth family converging uniformly to the payoff,

$$\varphi_{\varepsilon}(S) = \frac{1}{2}[(S - K) + \sqrt{((S - K)^2 + \varepsilon^2)}], \quad \text{with } |\varphi_{\varepsilon}(S) - V_0(S)| \leq \varepsilon/2, \quad (\text{C.25})$$

which is infinitely differentiable for $\varepsilon > 0$, so that the recursion (C.7) applies classically at every point including the strike, and the curvature discarded in Section C.5 is retained in finite form. The recursion is carried out in truncated Taylor (jet) arithmetic about the evaluation point, yielding the necessary derivatives with perfect accuracy. This is essential, not optional: noise in discretization is magnified by a factor of 10 for each derivative taken using finite differences, which breaks down the series after about three or four terms, whereas direct substitution of the nested radical (C.25) and subsequent symbolic expansion leads to an explosion in polynomial size that is not feasible beyond the second order.

The resulting series is not convergent but asymptotic. When $\varepsilon = 1.0$, the partial sums up to and including $N = 4$ become increasingly accurate, and the relative error at $N = 4$ is -0.07% , after which the accuracy starts to deteriorate, and at $N = 9$ the partial sums are an order of magnitude larger than the true price. This is the characteristic behaviour of an asymptotic expansion, such as the Stirling series or WKB expansions, which have an optimal order at which the series should be truncated; beyond this order, the accuracy of the approximation deteriorates. Based on this observation, we truncate our series at $N = 4$ and report the truncation order throughout.

Table C1 shows results from comparing across the α -cuts of a triangular fuzzy volatility with core $\sigma = 0.20$ and support $[0.17, 0.24]$ at $S = 110$, $K = 100$, $r = 0.03$ and $\tau = 1/12$, all within the scope conditions of Section 4. At each α this volatility is a closed interval, and we conduct the semi-analytic price at each endpoint independently. We include results at both endpoints since it is these endpoints that propagate through the pricing operator to form the price band, so any coding error at one endpoint would go undetected otherwise.

Table C1. Semi-analytic (ADM/HPM, regularized, $N = 4$, $\varepsilon = 1.0$) versus closed-form Black-Scholes prices at the endpoints of each α -cut of the fuzzy volatility number.

α	Endpoint	σ	Semi-analytic	Closed-form	Abs. error	Rel. error
0.00	lower	0.1700	10.303959	10.294140	0.009819	0.0954%
0.00	upper	0.2400	10.552747	10.508866	0.043881	0.4176%
0.05	lower	0.1715	10.305491	10.296666	0.008825	0.0857%
0.05	upper	0.2380	10.536680	10.500039	0.036641	0.3490%
0.10	lower	0.1730	10.307100	10.299279	0.007821	0.0759%
0.10	upper	0.2360	10.521463	10.491361	0.030102	0.2869%
0.50	lower	0.1850	10.323267	10.323414	-0.000147	-0.0014%
0.50	upper	0.2200	10.425872	10.427439	-0.001567	-0.0150%
0.90	lower	0.1970	10.347009	10.353418	-0.006410	-0.0619%
0.90	upper	0.2040	10.365550	10.373661	-0.008110	-0.0782%
1.00	core	0.2000	10.354471	10.361847	-0.007376	-0.0712%

Comparing across all α -cuts, the largest absolute relative difference occurs at the upper endpoint of the $\alpha = 0$ cut, where it is 0.42%, which is where the volatility is highest, and the regularization error is the largest in comparison with the value of the option (time value). It is less than 0.1% over the interior of the fuzzy number and is not larger than 0.02% for $\alpha = 0.50$. The error is dominated by the regularization parameter ε rather than the truncation order, and it decreases with ε at the cost of having to increase the truncation order to remain asymptotic.

We therefore consider the pricing implementation verified against the independent semi-analytic solution for all α -cuts utilized in the coverage tests. There are three caveats. Firstly, the verification is made for a single point in (S, K, r, τ) , as opposed to a grid of different moneyness and maturity values, and hence it is deemed to ascertain the correctness of the implementation at that representative point rather than across the entire panel. Secondly, recall that the regularized series is an asymptotic series; hence the reported agreement in Table C1 is an agreement at the optimal truncation and does not attest to the convergence of the regularized series as the number of terms grows

since it does not. Thirdly, the exact re-summation performed in Section C.7 yields the closed-form Black-Scholes price identically and hence serves to verify the derivation; it is the regularized calculation in this section that never refers to the closed form and that serves the purpose of verification.