

Session Vol Regime in SPY: A Walk-Forward Three-Way Classification

Martial Systems LLC | Working paper | 2026-09-18 | Not investment advice

Revisions. 2026-09-18: initial walk-forward diagnostics and research architecture freeze of multinomial logistic spec 00_logistic_regime_21d. Not live.

Diagnostics timestamp: 2026-09-18T13:09:52.879817+00:00. Out-of-sample window: 2019-01-02 to 2026-09-17 (n = 1,938). Research architecture freeze dated 2026-09-18. Related work: short-horizon realized-vol forecasts (production architecture frozen separately); session range expansion (research freeze, not live); open-to-close direction (rejected under pre-specified baselines).

1. Introduction

Write $R = (H - L) / O$ for the open-normalized session range of SPY. At the close of day t we ask which tercile of the trailing 252-session distribution of R will be realized on session $t+1$. The object is a three-way label (low, mid, high), not the sign of the open-to-close return and not the binary event that next-session R exceeds a ten-session mean.

Open-to-close sign prediction failed pre-specified accuracy and Brier criteria relative to a constant long baseline. Multi-day realized volatility is already in production under a separate freeze. Session range expansion, a binary large-versus-small classification against a short trailing mean, is a research freeze that is not live. The present exercise tests whether a coarser, year-relative regime of next-session range is distinguishable from climate and from simple lagged and volatility-proxy rules in walk-forward evaluation.

We report classification and probability-score diagnostics only. We do not evaluate option strategies or transaction costs of any trading rule. The production volatility specification is not modified here.

2. Data and sample design

Prices are daily open, high, low, close, and volume for SPY. When the production volatility repository already stores an OHLCV parquet, that file is copied; otherwise a public vendor API is used. Features observed at the close of day t are stored on the row dated $t+1$. An initial estimation sample ends 2018-12-31. Forecasts from 2019-01-02 onward are out of sample. Models are refit every 21 trading days on an expanding window of complete history strictly prior to the forecast origin. Shuffled cross-validation is not used.

In the reported diagnostics, the out-of-sample size is $n = 1,938$ with class rates low 0.357, mid 0.325, high 0.318.

3. Label and features

Let $q_{lo}(t)$ and $q_{hi}(t)$ be the one-third and two-thirds quantiles of R over the 252 sessions ending on day t . For prediction day $d = t+1$, the label is 0 if R_d is at most $q_{lo}(t)$, 1 if R_d is above $q_{lo}(t)$ and at most $q_{hi}(t)$, and 2 if R_d is above $q_{hi}(t)$. Ties at a boundary map down. This is not the ten-day mean expansion event and not a five-day realized-vol target.

The feature vector, all known at close t , includes lagged range (one day and a twenty-day mean), a twenty-day standardized range, close-to-close returns and annualized realized volatility at five and twenty days, a volume z-score, and day-of-week indicators for the forecast day. Same-day high, low, and close of $t+1$ are excluded. An ablation retains only five- and twenty-day realized volatility and a small range core (vol-proxy).

4. Models and baselines

The primary estimator is L2-regularized multinomial logistic regression with standardized inputs. A gradient-boosted tree is available in the same harness but is not the freeze specification. Probability vectors are compared to a climate forecast that assigns the out-of-sample class frequencies to every day. Hard-rule foils are constant mid (always class 1) and lagged regime (yesterday's realized tercile under the same quantile cut). Multiclass Brier is the mean squared error between the predicted class probabilities and a one-hot label.

Lagged-regime and constant-mid Brier scores treat those foils as one-hot probability vectors. That construction is harsh on hard rules. The tighter comparisons are zero-one accuracy versus lagged regime and Brier versus a fitted vol-proxy model that uses the same estimator on the ablation feature set.

5. Pre-specified evaluation criteria (2026-09-18)

Criteria were written in REGIME_DESIGN.md before the walk-forward peek. They are regime-native. They are not the always-up accuracy bar used for open-to-close direction, and they are not the ten-day expansion bars used for range expansion.

#	Criterion (primary)	Outcome
1	Multiclass Brier $\leq$ climate minus 0.002	satisfied
2	Improve on lagged regime (Brier minus 0.001 or acc plus 0.5 pp)	satisfied
3	Improve on constant mid (Brier minus 0.001)	satisfied
4	Full model improves on vol-proxy model (Brier minus 0.001 or acc plus 0.5 pp)	satisfied

Under these four primary criteria jointly, the logistic specification is recorded as **a research architecture freeze** (spec 00_logistic_regime_21d). That freeze authorizes documentation. It does not authorize live promotion of a regime model or any change to the production volatility architecture.

6. Out-of-sample results (2026-09-18)

Specification	Acc.	Brier	Climate Brier
Frozen logistic (full features)	0.5320	0.5677	0.6658
Logistic, vol-proxy features	0.5212	0.5889	0.6658
Lagged regime (hard)	0.5243	0.9515	0.6658
Constant mid	0.3251	1.3498	0.6658

Accuracy is 0.5320 against lagged regime 0.5243 (plus 0.77 percentage points; the bar was 0.5). Brier is 0.5677 against climate 0.6658 and against a fitted vol-proxy model 0.5889. Majority class is low (0.357); model accuracy 0.532 is not a live claim.

7. Accuracy by calendar year

Year	n	Accuracy	Brier
2019	252	0.5476	0.5691
2020	253	0.5573	0.5038
2021	252	0.5754	0.5628
2022	251	0.5219	0.5685
2023	250	0.4960	0.5983
2024	252	0.5000	0.6028
2025	250	0.5800	0.5278
2026	178	0.4551	0.6255

Yearly cells are descriptive. They are not additional freeze bars. Sample sizes per year are on the order of 250 trading days except the incomplete terminal year.

8. Limitations

- 1 Daily bars only; no residual range after the open on minute data.
- 2 No evaluation of options prices, implied volatility, or trading rules.
- 3 Vendor-adjusted OHLCV; microstructure and corporate-action conventions are not audited here.
- 4 Lagged-regime and constant-mid Brier treat hard rules as one-hot probabilities; those gaps overstate skill relative to accuracy versus lag.
- 5 A multi-day realized-vol production model is maintained separately and is not jointly estimated.
- 6 There is no paper hold-out stream of post-freeze closes yet.

9. Conclusion

Under the pre-specified primary criteria, next-session range tercile for SPY is statistically distinguishable from climate, from a constant-mid rule, from lagged regime, and from a fitted vol-proxy ablation in 2019 to 2026 walk-forward evaluation. The result is recorded as a research architecture freeze of multinomial logistic spec 00_logistic_regime_21d, not as a live product. This is not a recommendation to trade, and it does not alter the frozen multi-day realized-vol architecture. Session range expansion remains a separate research freeze. Open-to-close direction remains a rejected specification under its own criteria.

10. References (internal)

REGIME_DESIGN.md. FREEZE_DECISION.md. logs/regime_diagnostics.json. logs/walk_forward_summary.json.
logs/experiments/frozen_best_config.json. spy_vol_forecast methodology note and FREEZE_DECISION.md. spy_range_expansion
RANGE_DESIGN.md and FREEZE_DECISION.md. spy_direction_intraday FREEZE_DECISION.md. MODULE_EVAL_PATTERN.md.

Disclaimer

This document is for research documentation. It is not investment advice, not a solicitation to buy or sell any security or derivative, and not a performance record of live trading. Copyright (c) 2026 Martial Systems LLC. All rights reserved.