Conformal Prediction for Reliable Handover Under Distribution Shift

Johan Eliasson Eklund
https://github.com/elitan

***Abstract*—ML handover prediction is accurate in-distribution but fragile under shift. We evaluate conformal prediction (CP) for 5G handover under synthetic and real-world drift. We compare static CP, RAPS CP, Adaptive Conformal Inference (ACI), dynamic-step ACI (DACI), weighted CP, confidence-gated triggered ACI, and deep-ensemble CP against Top-k and a 3dB hysteresis baseline. On in-distribution synthetic data, static CP reaches 90.2% coverage. Under speed and noise shifts, RAPS reaches 95.7% and 94.2% coverage, while weighted CP reaches 91.1% in speed shift. Under severe shadow shift, static CP drops to 69.2%, RAPS improves to 79.6%, ACI restores 88.8%, and DACI reaches 89.9%. In regime-switch streams, DACI reaches 91.8%. On Irish 5G traces with speed-split drift, static CP reaches 76.4% coverage (95% CI $[66.8, 84.0]$), RAPS reaches 87.0% (95% CI $[77.2, 94.2]$), and DACI reaches 92.5% (95% CI $[89.9, 94.4]$). On Irish trace-holdout drift, DACI remains highest at 93.6%, while RAPS over-expands sets (size 70.3). Results show robust reliability under shift needs adaptive conformal control with explicit overhead budgeting.**

I. Introduction

Predictive handover reduces latency but can fail badly when radio conditions shift. Traditional 3GPP handover logic relies on hysteresis events [2], [3]. ML-based handover prediction improves point accuracy but provides no risk control when distribution changes [4], [12]. In production, that gap maps directly to radio link failure risk.

Conformal prediction (CP) gives finite-sample marginal coverage guarantees [5], [6]. Recent wireless CP work focuses on demodulation, channel tasks, and beam selection [7]–[9], while handover under distribution shift is still underexplored. To our knowledge, prior wireless CP literature does not report a handover-focused sequential-shift reliability study with real-trace validation.

Contributions.

1) We benchmark handover reliability under four synthetic shifts plus a regime-switch stream.
2) We compare static CP, RAPS CP, ACI [10], dynamic-step ACI, weighted CP, confidence-gated triggered ACI, and deep-ensemble CP under the same base predictor and KPI mapping.
3) We validate on Irish 5G driving traces with two protocols: source-target speed split and trace-holdout split.
4) We provide budget-aware reproducible runs (local-first, capped overflow policy) and release all generated artifacts.

II. Related Work

ML handover methods span supervised and reinforcement-learning policies [4], [12]. CP in wireless has shown value for calibration and reliability [9]. CP for beam selection shows strong reliability-efficiency tradeoffs [7], [8]. Appendix Table A1 and the released matrix file (`figures/related-work-matrix-v6.csv`) summarize scope differences across task, guarantees, shift handling, and real-trace validation. The key missing piece is still handover reliability under sequential shift.

III. System and Methods

A. Handover Prediction Setup

At time t, the model predicts future best cell $y_t = \arg\max_k \mathrm{RSRP}_k(t+H)$ using input

$$\mathbf{x}_t = [\mathrm{RSRP}_1, \ldots, \mathrm{RSRP}_K, \mathbf{e}_{c_t}, v_t]. \tag{1}$$

We use an MLP classifier and softmax scores $\hat{p}(y \mid \mathbf{x})$.

B. Baselines and Conformal Variants

3dB baseline: handover if best neighbor exceeds serving by 3dB.

Static CP:

$$\mathcal{C}(\mathbf{x}) = \{y : \hat{p}(y \mid \mathbf{x}) \geq 1 - \hat{q}\}, \tag{2}$$

with $\hat{q}$ calibrated on held-out source calibration data.

RAPS CP: regularized adaptive prediction sets using cumulative probability with rank penalty, calibrated on source and evaluated under shift.

ACI: online update of effective miscoverage level to track sequential drift [10].

DACI: dynamic-step ACI where the online step size switches between $(\gamma_{\text{low}}, \gamma_{\text{high}})$ using an EMA of recent error indicators.

Weighted CP: source calibration scores reweighted by estimated density ratio $w(\mathbf{x}) \propto p_T(\mathbf{x})/p_S(\mathbf{x})$ using a source-vs-target logistic discriminator.

Deep-ensemble CP: mean predictive probabilities from 5 independently seeded predictors, then static CP calibration on ensemble probabilities.

Triggered ACI: confidence-gated mixture that uses ACI sets when confidence is below a source-calibrated quantile threshold τ_q, otherwise static CP.

TABLE I
KEY DEFAULTS FOR REPRODUCIBILITY

Setting	Value
Miscoverage target α	0.10
Synthetic seeds	42, 123, 456, 789, 1011
RAPS defaults	$k_{reg} = 1$, $\lambda = 0.01$
DACI defaults	$\gamma_{low} = 0.005$, $\gamma_{high} = 0.02$, $\beta = 0.95$
Triggered-ACI default	confidence quantile $q = 0.7$
Deep-ensemble size	5 members
Irish speed-split traces	source 25, target 25
Irish trace-holdout traces	source 40, target 10

TABLE II
SYNTHETIC SHIFT COVERAGE (MEAN WITH 95% CI)

Shift	Top-1	Static CP	RAPS CP	ACI	DACI
IID	.677 [.658, .696]	.902 [.894, .910]	.958 [.953, .962]	.900 [.900, .900]	.944 [.944, .945]
Speed	.652 [.640, .665]	.897 [.886, .908]	.957 [.951, .962]	.900 [.900, .901]	.944 [.944, .945]
MeasNoise	.637 [.621, .652]	.880 [.872, .888]	.942 [.940, .944]	.900 [.899, .900]	.944 [.943, .944]
Shadow	.436 [.427, .444]	.692 [.686, .699]	.796 [.788, .803]	.888 [.885, .890]	.899 [.894, .904]
Regime	.531 [.512, .550]	.781 [.767, .796]	.863 [.854, .872]	.892 [.891, .893]	.918 [.915, .920]

C. System KPI Mapping

Coverage maps to handover success with bounded miss risk. Set size maps to measurement overhead. Undercoverage maps to RLF proxy. Serving-cell retention in small sets acts as implicit hysteresis and affects ping-pong rate.

IV. EXPERIMENTAL SETUP

A. Synthetic Shift Benchmark

Source setting: medium scenario (4×4 cells, $\sigma = 6$ dB shadowing, measurement noise 4 dB, speed 1–30 m/s, horizon $H = 10$). We train on source and calibrate on source only.

Target shifts:

1) IID (same as source)
2) Speed shift (20–50 m/s)
3) Measurement-noise shift (8 dB)
4) Shadow shift ($\sigma = 10$ dB)
5) Regime switch (source-like first half, harsh second half)

Each synthetic setting uses 5 seeds (42,123,456,789,1011), 600 trajectories/seed, and 20 training epochs. Main synthetic tables report means with 95% CIs.

B. Real-World Drift Benchmark

Dataset: Irish 5G driving traces [13]. Primary protocol uses source-target speed split: lower-speed 50% traces as source and higher-speed 50% traces as target, with source traces split 70/30 into train/cal. This gives 25 source and 25 target traces (36,238 train samples, 15,557 calibration, 30,398 target). Secondary protocol uses random trace-holdout (60/20/20 trace split): 40 source and 10 target traces (51,880 train, 17,957 calibration, 12,356 target). Model trains and calibrates on source only, then evaluates on target.

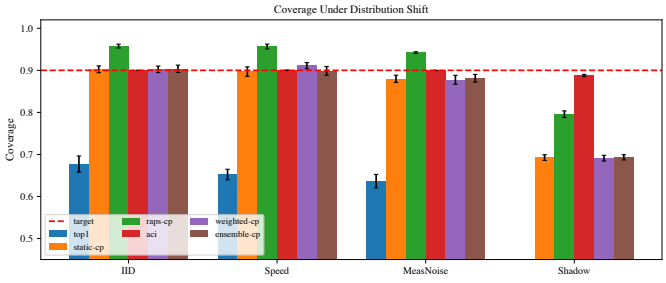

Fig. 1. Coverage across synthetic shifts. Static CP degrades under hard shift, RAPS improves moderate-to-hard shifts, and adaptive variants recover near-target reliability.

TABLE III
HARD-SHIFT RELIABILITY WITH UNCERTAINTY (MEAN OVER 5 SEEDS)

Shift	Method	Coverage (95% CI)	Avg Size	Overhead
Shadow	Static CP	.692 [.686, .699]	2.62	.170
	RAPS CP	.796 [.788, .803]	3.94	.314
	ACI	.888 [.885, .890]	5.98	.601
	DACI	**.899** [.894, .904]	6.35	.650
	Triggered ACI	.842 [.839, .844]	5.38	.541
	Weighted CP	.691 [.684, .698]	2.61	.169
	Ensemble CP	.693 [.687, .699]	2.61	.168
Regime	Static CP	.781 [.767, .796]	2.61	.168
	RAPS CP	.863 [.854, .872]	3.95	.307
	ACI	.892 [.891, .893]	4.50	.417
	DACI	**.918** [.915, .920]	5.04	.477
	Triggered ACI	.869 [.865, .874]	4.22	.387
	Weighted CP	.789 [.778, .799]	2.70	.176
	Ensemble CP	.784 [.771, .797]	2.60	.167

C. Default Settings and Split Sizes

V. RESULTS

A. Coverage Under Shift

Table II shows the main trend: static CP is reliable near source but degrades under strong shift (shadow, regime). RAPS gives clear gains over static under shift, and ACI/DACI keep coverage near target by expanding sets online.

B. Tradeoff in Hard Shifts

Paired seed deltas confirm hard-shift gains with uncertainty: RAPS vs static gives +10.3pp coverage on shadow (95% CI $[9.7, 10.9]$) and +8.1pp on regime (95% CI $[7.6, 8.6]$). ACI vs static gives +19.5pp and +11.1pp. DACI vs static gives +20.7pp and +13.6pp, and DACI vs ACI adds +1.1pp and +2.6pp. Ensemble CP is statistically close to static CP on hard shifts (+0.1pp shadow, +0.2pp regime). These paired CIs are estimated from 5-seed draws and should be read as finite-sample stability indicators.

C. Regime-Switch Stability

In regime-switch streams, static and weighted thresholds lag after the phase boundary. RAPS improves over static but still lags adaptive variants after abrupt transitions. ACI and DACI recover target-level coverage.

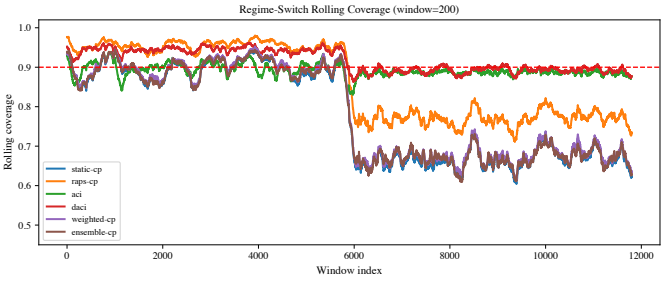

Fig. 2. Rolling coverage in regime-switch stream (window=200). ACI tracks the 90% target more closely than static and weighted CP.

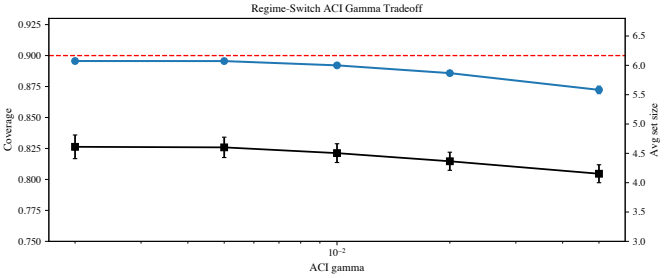

Fig. 3. Regime-switch ACI tradeoff over γ. Smaller γ improves coverage but increases set size and overhead.

D. ACI Step-Size Sensitivity

Figure 3 quantifies ACI sensitivity. In our regime-switch benchmark, small step sizes ($\gamma = 0.002$–0.005) achieve the highest coverage ($\approx 89.6\%$) with moderate set inflation, while larger values ($\gamma = 0.05$) reduce coverage to 87.2% but lower overhead. This supports tuning γ as a direct reliability-overhead control knob. Triggered ACI has a second control knob. A regime-switch trigger-quantile sweep in Appendix Fig. A1 shows quantile $0.5 \rightarrow 0.9$ improves coverage from 85.5% to 88.3% while overhead rises from 0.358 to 0.409. DACI hyperparameter sweeps over $(\gamma_{low}, \gamma_{high}, \beta)$ in Appendix Fig. A2 show a broad reliability-overhead frontier: low-step/high-reactivity settings reach up to 95.0% coverage in regime shift, while low-overhead settings stay near ACI coverage with smaller inflation. Conditional diagnostics in Appendix Fig. A3 and Fig. A4 show clear heterogeneity across speed and confidence deciles. On regime-switch speed deciles, worst-decile coverage rises from 65.8% (static) to 77.0% (RAPS), 88.2% (ACI), and 89.0% (DACI). On confidence deciles, worst-decile coverage rises from 74.0% (static) to 82.8% (RAPS) and 91.0% (DACI).

E. Real-World Drift Results

Size/133 is a normalized overhead proxy (average predicted set size divided by the 133 candidate cells in this dataset).

Trace-bootstrap on Irish speed-split confirms the reliability-cost pattern: RAPS vs static gives $+10.6$pp coverage (95% CI $[6.1, 15.8]$) with $+1.94$ set size (95% CI $[1.65, 2.30]$). ACI vs static gives $+11.6$pp (95% CI $[5.4, 19.4]$) with $+10.02$ set size (95% CI $[3.59, 18.84]$). DACI vs static gives $+16.1$pp (95% CI $[10.1, 23.7]$) with $+11.26$ set size (95% CI $[4.65, 20.17]$), and DACI vs ACI adds $+4.6$pp coverage (95% CI $[4.2, 4.9]$) with $+1.24$ set size (95% CI $[0.87, 1.70]$). Ensemble CP vs static stays flat in coverage ($+0.0$pp, 95% CI $[-0.3, 0.3]$) with slightly smaller sets (-0.10, 95% CI $[-0.14, -0.06]$). Appendix Fig. A5 compares ensemble CP against static and adaptive CP on shadow, regime, and Irish drift in one KPI view.

TABLE IV
IRISH 5G SPEED-SPLIT DRIFT RESULTS

Method	Coverage (95% CI)	Avg Size (95% CI)	Size/133
Top-1	.307 [.235, .379]	1.00 [1.00, 1.00]	.008
Top-3	.625 [.527, .710]	3.00 [3.00, 3.00]	.023
Static CP	.764 [.668, .840]	4.69 [4.45, 4.88]	.035
RAPS CP	.869 [.772, .942]	6.63 [6.43, 6.89]	.050
Weighted CP	.736 [.637, .816]	4.22 [4.00, 4.39]	.032
Ensemble CP	.764 [.667, .841]	4.59 [4.34, 4.79]	.035
Triggered ACI	.824 [.762, .872]	10.15 [6.23, 15.91]	.076
ACI	.879 [.856, .896]	14.70 [8.37, 23.44]	.111
DACI	**.925** [.899, .944]	15.94 [9.42, 24.87]	.120

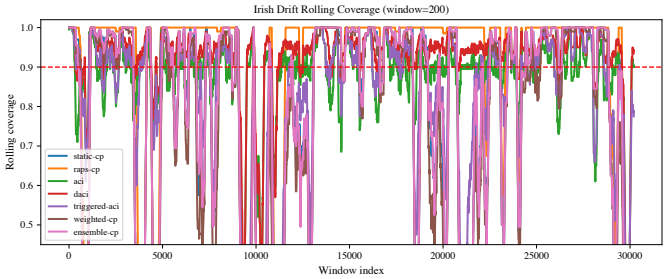

Fig. 4. Irish drift rolling coverage (window=200). DACI reaches highest reliability under source-target speed split.

F. Irish Trace-Holdout Stress Split

Trace-holdout is a stronger shift stress test. Here, RAPS vs static is negative in coverage (-1.4pp, 95% CI $[-3.1, -0.2]$) and very costly in set size ($+54.27$, 95% CI $[47.51, 59.77]$). This pattern is consistent with stronger support mismatch between source calibration and target traces, where RAPS rank-penalty regularization can over-expand sets. DACI remains strongest in coverage, but with higher overhead than static/weighted methods.

VI. DISCUSSION

When to use which method. Static or weighted CP is a low-overhead default for stable or mild shift. RAPS is a useful medium-budget upgrade when covariate shift is present and support overlap is still good. Triggered ACI is a practical middle point. ACI is robust for severe sequential drift. DACI is the max-reliability mode when overhead budget can absorb larger sets.

TABLE V
IRISH 5G TRACE-HOLDOUT DRIFT RESULTS

Method	Coverage (95% CI)	Avg Size (95% CI)	Size/133
Top-1	.153 [.077, .226]	1.00 [1.00, 1.00]	.008
Top-3	.661 [.521, .802]	3.00 [3.00, 3.00]	.023
Static CP	.896 [.782, .973]	16.07 [10.44, 22.97]	.121
RAPS CP	.882 [.752, .969]	70.34 [70.19, 70.49]	.529
Weighted CP	.889 [.769, .970]	15.07 [10.00, 21.25]	.113
Ensemble CP	.899 [.783, .975]	16.66 [10.85, 23.88]	.125
Triggered ACI	.909 [.825, .961]	16.43 [8.98, 26.08]	.124
ACI	.889 [.862, .904]	16.44 [7.17, 29.14]	.124
DACI	**.936** [.905, .953]	17.90 [8.17, 30.80]	.135

TABLE VI
MEASURED DEPLOYMENT POLICY UNDER OVERHEAD CAPS (SYNTHETIC HARD SHIFTS)

Budget Tier	Shadow Shift	Regime Switch
Low ($\leq$ .20)	Ensemble CP .693 [.687, .699] (.168)	Weighted CP .789 [.778, .799] (.176)
Medium ($\leq$ .40)	RAPS CP .796 [.788, .803] (.314)	Triggered ACI .869 [.865, .874] (.387)
High ($\leq$ .60)	Triggered ACI .842 [.839, .844] (.541)	DACI .918 [.915, .920] (.477)
Max ($\leq$ .70)	DACI .899 [.894, .904] (.650)	DACI .918 [.915, .920] (.477)

System implications. Reliability gains translate to lower RLF proxy but require explicit overhead budgeting. Table VI is measured from hard-shift runs and gives a direct policy map by overhead cap. The Irish trace-holdout stress split adds one warning: RAPS can over-expand sets under stronger support mismatch, so adaptive methods are safer for that regime. A coverage-overhead Pareto view in Appendix Fig. A6 shows triggered ACI and DACI both lie on the frontier for hard shifts, while static and weighted CP dominate low-overhead points. This policy table is synthetic-calibrated; deployment should re-tune thresholds and method switching on operator-specific traces before use.

A. Threats to Validity

Synthetic realism. Our synthetic channels capture controlled shift modes but simplify full multi-cell interference and scheduler behavior. **Irish feature scope.** Irish traces provide mobility and radio context, but not the full feature richness of operator measurement reports. **Split sensitivity.** Trace-holdout behaves differently from speed-split and exposes failure modes (for example RAPS set-size inflation), so deployment should validate split protocol assumptions. **Offline-online gap.** Results are offline replay; deployment needs online streaming integration, policy latency controls, and control-plane safety checks.

VII. CONCLUSION

We presented a shift-focused handover reliability study with conformal prediction. Static CP works well in-distribution but degrades in severe shift. RAPS improves over static in many synthetic and Irish speed-split settings, but fails under stronger trace-holdout mismatch due to large set inflation. ACI restores near-target coverage under shadow and regime-switch drift. DACI pushes reliability further in hard shifts and Irish splits, with additional overhead. Weighted CP improves moderate shifts with small inflation. Triggered ACI gives a practical middle tradeoff. The core practical result is clear: robust handover reliability needs adaptive conformal control plus explicit budget-aware policy selection.

TABLE A1
RELATED-WORK EVIDENCE MATRIX

Work	CP	Shift	Real	Handover
Lee [12]	No	No	Yes	Yes
Yajnanarayana [4]	No	No	Yes	Yes
Cohen [9]	Yes	No	Partial	No
Hegde [7]	Yes	Limited	No	No
Deng [8]	Yes	Limited	Yes	No
This work	Yes	Yes	Yes	Yes

TABLE A2
APPENDIX: STATIC VS RAPS VS ENSEMBLE (V6)

Scenario	Static CP	RAPS CP	Ensemble CP
Shadow (synthetic)	.692 [.686, .699]	.796 [.788, .803]	.693 [.687, .699]
Regime (synthetic)	.781 [.767, .796]	.863 [.854, .872]	.784 [.771, .797]
Irish (speed-split)	.764 [.668, .840]	.869 [.772, .942]	.764 [.667, .841]

APPENDIX A
APPENDIX: RELATED WORK, DIAGNOSTICS, ENSEMBLE AND LATENCY

Measured medium-scenario latency: calibration 0.08 ms, CP set construction $1.59\,\mu$s/sample, NN inference $0.50\,\mu$s/sample.

REFERENCES

[1] T. S. Rappaport *et al.*, "Millimeter wave mobile communications for 5G cellular," *IEEE Access*, 2013.
[2] 3GPP TS 38.214, "NR; Physical layer procedures for data," 2022.
[3] 3GPP TS 38.331, "NR; Radio Resource Control (RRC); Protocol specification," 2022.
[4] V. Yajnanarayana, H. Rydén, and L. Hévizi, "5G handover using reinforcement learning," in *Proc. IEEE 5GWF*, 2020.
[5] V. Vovk, A. Gammerman, and G. Shafer, *Algorithmic Learning in a Random World*. Springer, 2005.
[6] A. N. Angelopoulos and S. Bates, "A gentle introduction to conformal prediction," *arXiv:2107.07511*, 2021.
[7] D. N. Hegde *et al.*, "Reliable and efficient beam selection using conformal prediction in 6G systems," in *Proc. IEEE WCM*, 2024.
[8] J. Deng *et al.*, "SCAN-BEST: Sub-6GHz-aided near-field beam selection with formal reliability guarantees," *arXiv:2503.13801*, 2025.
[9] K. M. Cohen *et al.*, "Calibrating AI models for wireless communications via conformal prediction," *IEEE TMLCN*, 2022.
[10] I. Gibbs and E. Candès, "Adaptive conformal inference under distribution shift," *NeurIPS*, 2021.
[11] Y. Romano *et al.*, "With malice toward none: Assessing uncertainty via equalized coverage," *HDSR*, 2020.
[12] W. Lee *et al.*, "Prediction-based conditional handover for 5G mm-wave networks," *IEEE Access*, 2020.
[13] D. Raca *et al.*, "Beyond throughput: A 5G dataset with channel and context metrics," in *Proc. MMSys*, 2020.
[14] B. Lakshminarayanan *et al.*, "Simple and scalable predictive uncertainty estimation using deep ensembles," in *NeurIPS*, 2017.

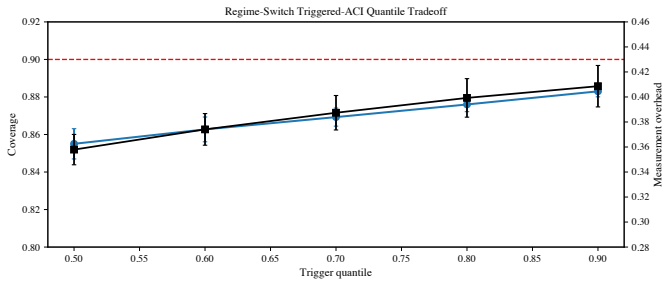

Fig. A1. Regime-switch triggered-ACI tradeoff over confidence-gate quantile. Higher quantiles increase coverage and overhead.

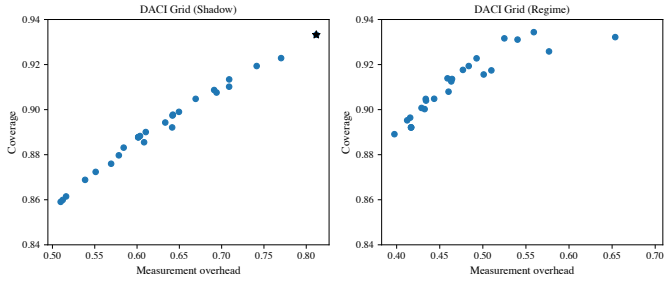

Fig. A2. DACI robustness grid on shadow and regime shifts. Each point is one $(\gamma_{low}, \gamma_{high}, \beta)$ setting.

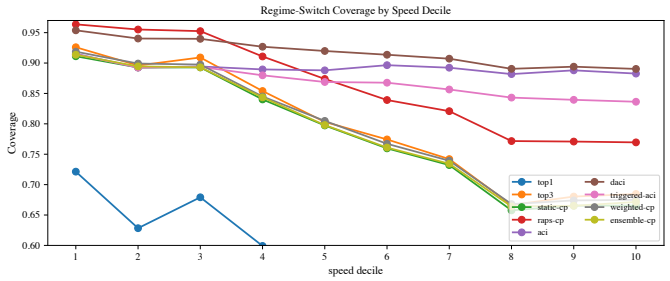

Fig. A3. Synthetic regime-switch coverage by speed decile across methods.

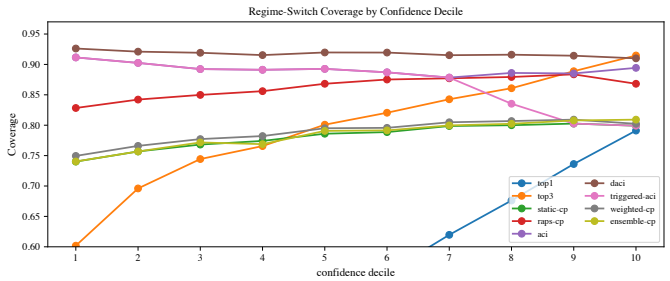

Fig. A4. Synthetic regime-switch coverage by confidence decile across methods.

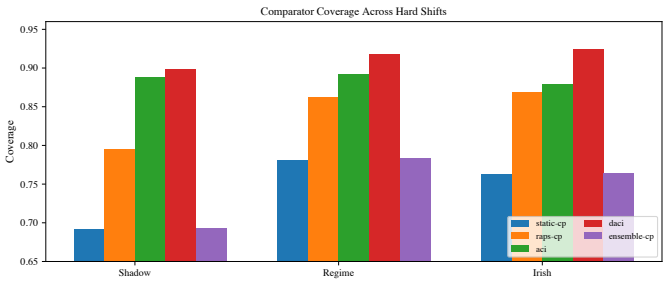

Fig. A5. Comparator view: ensemble CP vs static and adaptive CP across shadow, regime, and Irish drift.

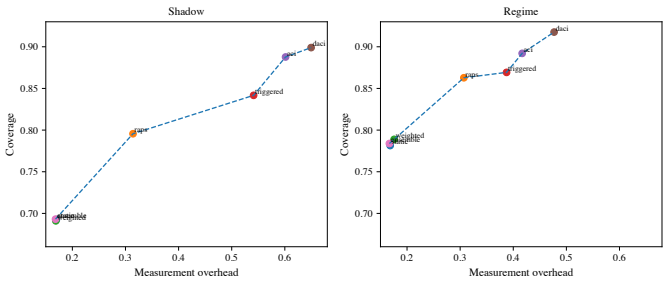

Fig. A6. Hard-shift Pareto map (coverage vs overhead). Triggered ACI and DACI form practical frontier points for medium and high overhead budgets.