Online Resource 3 Additional source confirmation

Label Budget Trade-offs and Calibration Transfer in GNSS Spoofing Detection

Annals of Telecommunications
Bo Xie and Yi Zhao
Wuhan Digital Engineering Institute, Wuhan, China
Corresponding author: Bo Xie (xiebo@mail.hfut.edu.cn)

Design and scope

The additional test used the published Blacksburg Sep28 G4/S4 files with PRN 9, selected by format compatibility and the smallest common directory PRN before inspecting model performance. Previous contact with other files in this cohort was limited to source, format and content auditing. Exactly 2,000 rows per recording condition were selected using the original deterministic hash-jittered source-index stratification. During this confirmation, the 4,000-row test set was not used for fitting, scaling, threshold estimation or model selection. Subsequently it became part of the exposed development pool described in Online Resource 5.

All 135 archived LR, RF and metric checkpoints and all 3,240 original threshold cells were reused. Six Arlington training/calibration role assignments and five support seeds were retained. These are reused experimental conditions on one additional test cohort, not independent acquisitions. The near-calibration arm is repeated across the two held-out-role assignments for a given training source. All conditions and failures are reported separately from the original experiment.

This is within-dataset confirmation under a joint change in genuine-recording location, spoof generation mechanism, PRN and acquisition background. It is neither independent-dataset validation nor an isolated causal test of location. Labels denote published recording conditions, not verified receiver takeover at every observation. No physical-time or detection-delay claim is made.

The initial local protocol was retained with a documented numerical implementation amendment after the first scoring attempt was rejected by validation. Float32 score storage incorrectly differed from the archived float64 scoring convention, and scalar promotion rounded thresholds during comparison. The accepted run preserves float64 scores and the original strict comparison to the unchanged thresholds. Samples, models, calibration banks, conditions and reporting scope did not change. The amendment affected numerical implementation only; the protocol was locally frozen rather than publicly preregistered.

Main confirmation results

The table uses k=20, c=160, nominal alpha=0.05 and held-out normal calibration. New results are means over the six source-role assignments and five seeds. Old results use the exactly matched original cells; they equal the old source-macro mean under the balanced design. The new-minus-old difference describes different test distributions and is not a causal effect of a single factor.

Table OR3.1 is provided as Table_OR3_1.csv: 3 data rows. Column headings and displayed precision are preserved from the archived supplement. The table is not filtered.

RF and metric did not retain their original discrimination on this additional source. Their directed-score AUROCs were far below 0.5, while their false-negative rates approached 100%. The prespecified direction was retained: no score reversal or test-dependent recalibration was performed. LR had lower errors than on the old sources at this operating point, but its mean AUROC remained approximately 0.504. Thus, a change in thresholded error rates alone does not establish a generally useful ranking.

For fixed B=200, increasing k from 5 to 50 while reducing c from 190 to 100 increased RF FPR by 11.578 percentage points and FNR by 0.232 points. For metric, FPR increased by 19.927 points while FNR decreased by only 0.053 points. The large miss-rate improvement seen on the original sources did not transfer to this added condition. This is a result for the frozen allocation contrast, not an impossibility claim for other training or adaptation strategies.

At k=20, increasing held-out normal calibration from c=20 to c=500 reduced RF and metric FPR by 11.905 and 6.158 points, respectively, but did not restore useful attack detection: FNR remained near 100%. At the intermediate allocation, switching from near to held-out calibration reduced RF FPR by 13.495 points, with only a 0.025-point FNR reduction. Calibration may move an operating point without recovering the score ordering needed in the new condition.

Verification and sensitivity

The independent validator checked source import, frozen sample identities, all checkpoint states, all 135 new score vectors, all 270 archived test-score replays, all original calibration thresholds and all 3,240 metric rows. Maximum replay error was zero for both new and archived scores. Three resealed single-fault copies with an incorrect false-positive count, threshold or source label were rejected.

The primary test retained 12 genuine and 2 counterfeit CN0-zero rows. The prespecified fixed-score sensitivity excluded only these test rows, without refitting or recalibrating. Across all cells, the largest absolute changes were 0.2440 FPR points and 0.0642 FNR points. This limited check does not explain the large degradation, but does not rule out other acquisition or processing effects.

These findings narrow the scope of the original allocation results. They support a conditional interpretation: budget allocation effects depend on whether the frozen scoring model remains discriminative under the target conditions. They do not identify the individual cause of failure, establish population error guarantees, or constitute a new detector. No p-values or population confidence intervals are reported.

Complete condition means

All 108 means are supplied in the accompanying table; the 648 role-specific means and 2,700 raw matched comparisons remain in the accompanying CSV files. Rates are percentages and AUROC is unitless. Repeated role/seed cells are not counted as independent test cohorts.

Table OR3.2 is provided as Table_OR3_2.csv: 108 data rows. Column headings and displayed precision are preserved from the archived supplement. The table is not filtered.