Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Galactic-Latitude Dependence of the Blend-Covariate Coefficient in Proper-Motion Contamination Tests: A Fifteen-Field Analysis of unTimely Epoch Astrometry

Ritwik Kumar

ritwik.ricko@gmail.com

Article

Keywords:

Posted Date: November 27th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11723707/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Galactic-Latitude Dependence of the Blend-Covariate Coefficient in Proper-Motion Contamination Tests: A Fifteen-Field Analysis of unTimely Epoch Astrometry

Ritwik Kumar
Independent Researcher
Noida, Uttar Pradesh, India[⋆]

ABSTRACT

Meisner et al. (2021), reading a figure in Marocco et al. (2021), infer that CatWISE2020 proper-motion uncertainties near the south ecliptic pole at $W1 \approx 17.2$ are underestimated by $\approx 2.1\times$. We could not re-derive this; it motivates our analysis but is not an input to it. Using unTimely epoch-level astrometry (from time-resolved unWISE coadds like CatWISE2020's, but measured with different software), we test whether excess epoch-level scatter tracks an independent blending indicator, `fracflux`. A protocol written before the main analysis was applied to fifteen fields spanning $|\mathrm{BGAL}| = 2.2$–$77.5°$ (222,828 sources; 5.68×10^6 epoch records); the last field was added afterwards as an out-of-sample test. The protocol's confirmation criterion fails: the `fracflux` coefficient has no consistent sign across fields (Cochran's $Q = 1276.5$ on 14 d.f.; $I^2 = 98.9\%$). A post-hoc meta-regression shows it rising toward the Galactic plane ($\hat{\beta}_1 = 1.158 - 0.0370\,|\mathrm{BGAL}|$; slope $p = 8 \times 10^{-4}$ after scaling for residual heterogeneity), and the slope is negative in all eight model specifications tried. Whether the coefficient changes sign, and at what latitude, depends on the specification (zero crossing 0.2–31.3°); Fields 3 and 8 ($|\mathrm{BGAL}| = 2.7°$, $8.7°$) are significantly reversed in all eight. Leave-one-out prediction gets the sign right in 14 of 15 fields for the executed model, and Field 15's sign was predicted from the other fields before it was processed (on our own record). Two candidate physical accounts we tested are insufficient on their own. A latitude-calibrated score is validated only on these fields. Independent replication, and tests on CatWISE2020's own epoch measurements, remain open.

Key words: astrometry – methods: statistical – methods: data analysis – proper motions – surveys – stars: low-mass

1 INTRODUCTION

Citizen-science motion selection, most notably the Backyard Worlds: Planet 9 project (Kuchner et al. 2017), finds moving sources by visual inspection: volunteers view animated "flipbooks" of difference images built from time-resolved WISE coadds. The project's first published discovery, the T5.5 dwarf WISEA J110125.95+540052.8, was found 0.9 mag fainter than the $W2$ single-exposure sensitivity (Kuchner et al. 2017). Motion-selected discoveries since then include three extreme-T-subdwarf candidates (one previously reported) with total proper motions of 0.9–2.2″ yr^{-1} (Meisner et al. 2021) and, found by Backyard Worlds volunteers, the hypervelocity L subdwarf CWISE J124909.08+362116.0, whose Galactic-rest-frame speed is $456 \pm 27\,\mathrm{km\,s^{-1}}$ and for which fewer than 0.002% of Gaia stars within 200 pc have comparable tangential velocities (Burgasser et al. 2024).

Selecting on a large measured motion favours objects whose measured motions are large for any reason, including measurement error, so the reliability of quoted proper-motion uncertainties matters for such searches. Discovery papers acknowledge the selection effect: Meisner et al. (2021), discussing WISEA J155349.96+693355.2 ($\mu = 2157 \pm 55\,\mathrm{mas\,yr^{-1}}$; $V_{\rm tan} \approx 395\,\mathrm{km\,s^{-1}}$ from a photometric distance), write that "It would be surprising if the WISEA 1553+6933 $V_{\rm tan}$ were truly so high as $\sim 400\,\mathrm{km\,s^{-1}}$" and that "our motion selection methodology biases us toward finding objects with preferentially high $V_{\rm tan}$". Whether, and where on the sky, a catalogue's quoted proper-motion uncertainties understate the real scatter is the practical question behind the present work.

Motion catalogs derived from WISE contain many spurious high-proper-motion entries. Two recent examples: among 11,900 automatically detected candidates, 10,170 were identified as WISE artifacts, diffraction spikes or ghosts during visual vetting (Brooks et al. 2024); and a CatWISE2020-based candidate list for the Backyard Worlds: Cool Neighbors project has an expected efficiency of only $\sim 0.5\%$ (Karpov et al. 2025). Selection cuts on blend indicators are a common mitigation: Karpov et al. (2025), for example, require `fracflux` > 0.5 "to ensure that the detections are not subject to significant crowding". A fixed threshold of this kind

⋆ ritwik.ricko@gmail.com

© 2026 The Authors

implicitly assumes that the indicator means the same thing in every part of the sky. This paper tests that assumption. We rule out simpler mechanisms analytically (Section 4), build and validate the statistical machinery (Sections 4–6), execute it on real data at the scale the protocol set (Section 7), and then test whether the resulting single-covariate diagnostic is stable across the sky (Section 7.3). It is not; we show how it fails, how robustly, and in which respects the failure is not a statistical artefact of our own model.

2 SCIENTIFIC BACKGROUND

WISE (Wright et al. 2010) surveyed the sky at $3.4, 4.6, 12, 22\,\mu$m ($W1$–$W4$) from January 2010; after cryogen exhaustion the mission continued as NEOWISE, then NEOWISE-Reactivation (NEOWISE-R) from December 2013, revisiting the sky roughly twice yearly. CatWISE2020 (Marocco et al. 2021), the successor to the CatWISE Preliminary Catalog (Eisenhardt et al. 2020), contains 1,890,715,640 sources and uses WISE/NEOWISE data from 2010 January 7 to 2018 December 13 (≈ 8.9 yr). Its positions and motions are measured with an adapted version of the AllWISE `WPHOT` software on 12 time-resolved unWISE epoch coadds, with a `crowdsource`-derived unWISE catalog as the detection list (Marocco et al. 2021, Sections 2.1–2.3). The authors report proper-motion accuracies (scatter against Gaia DR2) of $\sim 20\,\mathrm{mas\,yr^{-1}}$ at $W1 \sim 15$ and $\sim 100\,\mathrm{mas\,yr^{-1}}$ at $W1 \sim 17$, a 5σ limit of $W1 = 17.43$ (Marocco et al. 2021, Section 3.2), and 90% completeness at $W1 = 17.7$. Positions and motions are *not* corrected to the Gaia frame: tile-dependent offsets of typically ± 150 mas (position) and $\pm 40\,\mathrm{mas\,yr^{-1}}$ (motion) remain, and the authors recommend correcting them tile by tile with their Table 1 (Marocco et al. 2021, Section 3.3.1).

unTimely (Meisner et al. 2023) provides `crowdsource` (Schlafly et al. 2019) photometry and astrometry on time-resolved unWISE coadds: 16 coadd epochs per band at a typical sky location, at least 15 everywhere and up to 295 near the ecliptic poles, with a typical baseline of 10.5 yr (early-to-mid 2010 to mid-to-late 2020; NEOWISE through 2020 December) (Meisner et al. 2023, Section 2). WISE's scanning geometry gives far deeper coverage near the ecliptic poles than at low ecliptic latitude. unTimely astrometry is *not* recalibrated to Gaia, and its authors caution that astrometric systematics can reach ~ 250 mas per coordinate (Meisner et al. 2023, Section 8). unTimely is therefore a closely related but not identical measurement to CatWISE2020's: both use time-resolved unWISE coadds, but with different measurement software (`crowdsource` versus `WPHOT`), a different number of epochs (16 versus 12) and a different baseline. That the epoch-level behaviour measured here in unTimely carries over to CatWISE2020's own epoch-level measurements, which are not published, is an assumption of this paper, not a result (Section 12).

The motivating benchmark, and what it rests on. Meisner et al. (2021, Section 4.1) compare CatWISE2020 and Gaia DR2 proper motions using Figure 23 of Marocco et al. (2021), which covers the south-ecliptic-pole tile (0890m667), and state that near the pole at $W1 = 17.2$ the median χ^2 is ~ 2 for both μ_α and μ_δ, against 0.45 expected for one degree of freedom if the uncertainties were correct; they conclude that the quoted uncertainties are underestimated there by $\sqrt{2/0.45} = 2.1$. We have not found this number stated in the text of Marocco et al. (2021): it is read off a plot, so it is approximate (we write ≈ 2.1 and give no further digits), and it is specific to one sky region and magnitude – for a different object near the ecliptic plane, Meisner et al. (2021, Section 4.3) infer an underestimate of only $\sim 7\%$ from Figure 13 of Marocco et al. (2021). Marocco et al. (2021) themselves attribute the generally worse motion χ^2 they see against Gaia, including at the poles, to the lowered motion-uncertainty floor (their Sections 2.3 and 3.3.2). We could not reproduce the figure ourselves, because CatWISE2020–Gaia cross-matches are not reachable from our computing environment. The benchmark is therefore motivation only: no result below uses it as an input or a fitting target, and none of our conclusions depends on it being exactly right.

3 LITERATURE REVIEW AND RESEARCH GAP

Marocco et al. (2021) give one quantified internal reason for worse motion χ^2 against Gaia: the motion-uncertainty floor was lowered from $10\,\mathrm{mas\,yr^{-1}}$ in the Preliminary Catalog (Eisenhardt et al. 2020, Appendix B) to $1\,\mathrm{mas\,yr^{-1}}$ in CatWISE2020, and the new floor is "significantly smaller than the minimum measured scatter with respect to Gaia motion" (their Appendix A). We show (Section 4.3) that a floor of this size cannot account for a factor of ≈ 2 at $W1 \approx 17.2$, where the quoted accuracy exceeds either floor by about an order of magnitude or more. They also note that "the astrometry for faint sources is more susceptible to degradation due to blending" (Section 3.3.1) and that fast-moving sources are "smeared" in the full-depth coadds used for detection, producing multiple detections (Section 3.3.3).

Other work reports high rates of spurious motions in WISE-based searches, without a model of how they depend on position on the sky. Brooks et al. (2024) identified 10,170 of 11,900 SMDET candidates as WISE artifacts, diffraction spikes or ghosts during visual vetting (their Section 2.2) – a different failure mode from blending, with the same practical burden. Karpov et al. (2025) select on the `fracflux` indicator used in this paper, requiring `fracflux` > 0.5 "to ensure that the detections are not subject to significant crowding" (their Section 2); they attribute spurious cross-match associations with CatWISE2020 to "systematic deblending problems in the catalogue" (Section 2.2), and quote an expected efficiency of about 0.5% for the CatWISE2020-based input list of the Backyard Worlds: Cool Neighbors project, "due to large amount of spurious high proper motion entries" (Section 3). In the literature we reviewed we found no test of whether a blend indicator's relation to astrometric error is the same in different parts of the sky; this paper makes that test for one indicator in one survey, and says nothing about what the works above would have found had they done the same.

Heteroscedasticity-consistent covariance estimators (White 1980) are biased downward in small samples because squared residuals underestimate the error variance; MacKinnon & White (1985) introduced leverage-based corrections. We apply the same idea to each source's motion fit (leverage-

corrected studentized residuals; Section 4), where only $K \sim$ 6–17 epochs are available.

The gap. In the literature reviewed above we found no derivation of how a stochastic epoch count in a threshold-limited fit propagates into miscalibration (Section 4 derives it and finds the effect negligible here); no quantified contamination model tied to an independent blending covariate with a pre-specified multi-field test; and no test of whether such a covariate's inferred effect is stable across the sky. These are statements about what we found, not claims about the whole literature.

4 THEORETICAL FRAMEWORK

4.1 Leverage-corrected studentized residuals

For $y_i = a + bt_i + e_i$, $e_i \sim \mathcal{N}(0, \sigma_i^2)$, weights $w_i = 1/\sigma_i^2$, the normal equations

$$a\sum w_i + b\sum w_i t_i = \sum w_i y_i, \tag{1}$$

$$a\sum w_i t_i + b\sum w_i t_i^2 = \sum w_i t_i y_i \tag{2}$$

give, with $S_w = \sum w_i$, $S_{wt} = \sum w_i t_i$, $S_{wt^2} = \sum w_i t_i^2$, $D = S_w S_{wt^2} - S_{wt}^2$,

$$\hat{b} = \frac{S_w S_{wty} - S_{wt} S_{wy}}{D}, \quad \hat{a} = \frac{S_{wy} - \hat{b} S_{wt}}{S_w}. \tag{3}$$

In the whitened design $Z_i = \sqrt{w_i}[1, t_i]$, this is OLS with hat matrix $H^* = Z(Z^\top Z)^{-1} Z^\top$; direct computation gives

$$h_i = \frac{w_i(S_{wt^2} - 2t_i S_{wt} + t_i^2 S_w)}{D}, \quad \sum_i h_i = 2, \tag{4}$$

verified against brute-force $Z(Z^\top Z)^{-1}Z^\top$ to 1.1×10^{-16}. Since the naive normalized residual e_i/σ_i has variance $(1 - h_i) < 1$ under H_0 – an artifactual *deflation* – a direct simulation (7 irregular epochs) gives uncorrected variance $(0.607, 0.718, 0.808, 0.855, 0.833, 0.707, 0.469)$ across epoch position, while

$$z_i \equiv \frac{e_i}{\sigma_i \sqrt{1 - h_i}} \tag{5}$$

gives $(1.000, 1.000, 1.000, 1.001, 0.996, 0.996, 1.003)$ – unity everywhere to simulation precision. All analysis below uses z_i.

4.2 Epoch-count stochasticity: derived and ruled out

Proposition 1. For $\hat{\mu} \mid K \sim \mathcal{N}(\mu_{\rm true}, C/K)$, K random, the marginal excess kurtosis is $3\,\mathrm{Var}(K^{-1})/E(K^{-1})^2 \geq 0$, zero only for deterministic K (full derivation, Appendix A).

With $K \sim \mathrm{Binomial}(N_{\max}, \bar{p}(W1)) \mid K \geq K_{\min}$ and literature parameters ($N_{\max}$=12, $K_{\min}$=3, $W1_{5\sigma}$=17.67), direct Monte Carlo of the full WLS fit (40,000 simulated sources per point, seeded; Figure 1) gives a factor between 0.986 and 1.012 across $W1$=14.5–18.2 (single-point Monte Carlo standard deviation at most 0.009, from 20 independent replicates per point; the 20-replicate means lie between 0.9980 and 1.0023, within 2.0 standard errors of 1 at every point), including below the detection limit. This is expected analytically: with uniform formal weights the standardized slope is exactly $\mathcal{N}(0,1)$ given each realized K, hence also marginally, so epoch-count stochasticity changes the tails of the *unstandardized* slope but not the calibration of the formal uncertainty.

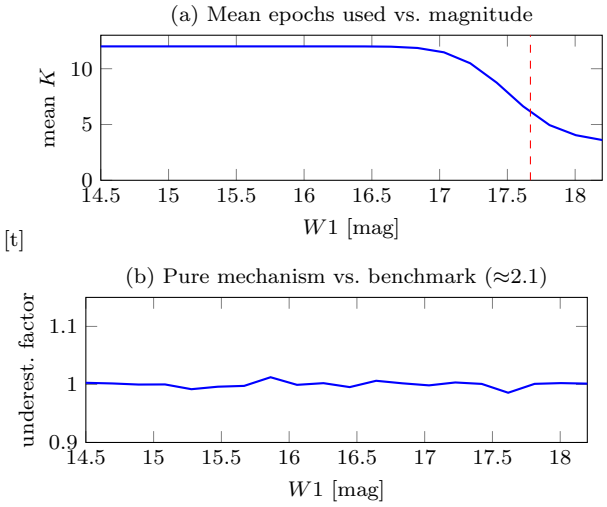


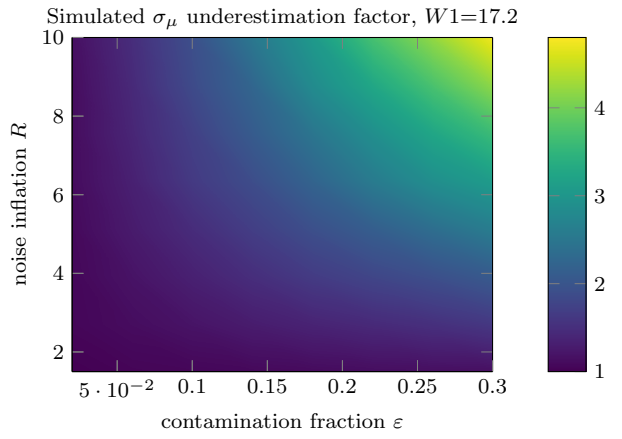
Figure 1. Pure epoch-count-stochasticity mechanism, direct Monte Carlo (40,000 trials per point), literature-anchored parameters only. The predicted factor is consistent with 1 at every point (largest excess over unity 0.012, against the ≈ 1.1 excess of the ≈ 2.1× benchmark).

4.3 Simple alternatives excluded: H_1 and H_2

H_1 (uniform rescaling f): taking the benchmark median $\chi^2 = 2$ as exact, $f = \sqrt{2/0.4549} = 2.097$ reproduces it by construction; the discriminating question is tail realism (below). H_2 (fixed additive floor c): the predicted factor is $\sqrt{1 + c^2/\sigma_{\rm formal}^2}$. We take the scatter against Gaia quoted by Marocco et al. (2021), $\sim 20\,\mathrm{mas\,yr^{-1}}$ at $W1 \sim 15$ and $\sim 100\,\mathrm{mas\,yr^{-1}}$ at $W1 \sim 17$, interpolated log-linearly to $117.5\,\mathrm{mas\,yr^{-1}}$ at $W1 = 17.2$, as the true uncertainty; if the formal uncertainty is smaller by the benchmark factor, $\sigma_{\rm formal} \approx 117.5/2.097 = 56.0\,\mathrm{mas\,yr^{-1}}$ and the benchmark requires $c = \sigma_{\rm formal}\sqrt{2.097^2 - 1} \approx 103\,\mathrm{mas\,yr^{-1}}$. The largest additive systematic quoted by Marocco et al. (2021) (the typical tile-level motion offset, $c = 40\,\mathrm{mas\,yr^{-1}}$; their Section 3.3.1) gives a factor of only 1.23, and the required c is 2.6 times larger. H_2 is therefore disfavoured. It is the less secure of the two exclusions, because $\sigma_{\rm formal}$ is interpolated from quoted accuracies rather than read from the catalog.

4.4 Contamination mixture and identifiability

A fraction ε of epochs contaminated, true noise inflated by R, pipeline unaware (uniform assumed weight). Figure 2 shows the resulting factor over (ε, R) at $W1$=17.2 (25,000 trials/point). The benchmark factor $f = \sqrt{2/0.4549} = 2.097$ is reached inside the plotted range along a continuous, degenerate contour, not a unique point (e.g. ε≈0.18, R≈5.9 or ε≈0.30, R≈4.0; the contour exists for $\varepsilon \geq 0.14$ on this grid). Re-simulating 5 contour members with 400,000 sources each, every one has heavier tails than H_1 at the same median factor (excess kurtosis 0.8–2.5; $P(|z|{>}5)$ larger than under H_1 by a



[t]

Figure 2. Contamination-mixture parameter space, direct Monte Carlo. The $\approx 2.1\times$ benchmark is reproduced along a continuous, degenerate contour, not a unique (ε, R).

factor of 2.4–5.9), a falsifiable signature H_1 cannot mimic even though (ε, R) remain unidentified individually.

4.5 Hypothesis hierarchy

H_0: correctly calibrated Gaussian epoch errors;
H_1: global multiplicative rescaling (benchmark-reproducing, tail-inconsistent);
H_2: fixed additive floor (disfavoured, §4.3);
H_3: generic heavy-tailed errors, no covariate;
H_4: contamination mixture, no covariate (β_1=β_2=β_3=0);
H_5: contamination mixture depending on `fracflux`;
H_6: false cross-epoch association, screened *before* residual inference.

A recovery test (simulate true H_3, fit all models), recorded in the pre-registration (`PREREGISTRATION.md`, Sections 9 and 4b), found H_3 and H_4 not reliably distinguishable from residual shape alone; the numerical output of that test was not retained, so we quote no figures for it. We route the decisive test through H_5's external covariate instead of forcing this resolution.

4.6 Marginal model with cluster-robust inference

Epochs within a source share fitted parameters; we adopt a marginal logistic model with the Liang–Zeger sandwich (Liang & Zeger 1986):

$$V = A^{-1}\Big[\sum_j \big(\sum_{i\in j} s_i\big)\big(\sum_{i\in j} s_i\big)^\top\Big]A^{-1} \tag{6}$$

(full derivation, Appendix B). A synthetic validation with realistic mixed between/within-source `fracflux` structure gives naive coverage 75.3% under clustering, restored to 92.7% by Eq. 6; a finite-cluster correction helps but does not fully resolve small-G coverage (G=8: 85.8% → 90.0%), so cross-field checks throughout this paper use per-field separate estimates, never a single pooled fit with field as an additional small-G cluster level.

4.7 Model comparison: QIC

AIC's $2k$ penalty assumes independence; we use (Pan 2001)

$$\mathrm{QIC} = -2\hat{L} + 2\,\mathrm{tr}(HV_{\mathrm{robust}}), \tag{7}$$

reducing to AIC when $V_{\mathrm{robust}} = A^{-1}$. In the simulation audit of Appendix K (N=10,000 sources; 300 simulated datasets per design point) the cluster-robust Wald test of β_1 at $p < 0.01$ had power at least equal to that of the literal 3-d.f. ΔQIC$>$ 10 comparison in 17 of the 19 design points with $\beta_1 > 0$ (largest advantage 25.3 percentage points, at K=4, β_1=0.3, σ_u=1: 78.3% against 53.0%); the 2 exceptions favoured ΔQIC by at most 6.3 points (at K=9, β_1=0.3, σ_u=1.5), while both rules had Type-I rates of at most 1.3%. This is why the protocol makes the Wald test primary and ΔQIC corroborating. We use QIC rather than AIC because the records of one source are not independent (Pan 2001); that choice rests on this argument, not on a simulation of ours.

4.8 Natural cubic splines

$W1, K$ enter via natural splines (Hastie et al. 2009) (full construction, Appendix C), 2 boundary + 1 interior knot, fit once and reused identically (verified by direct equality test). Degenerate knots (found on real data, confirmed again below for four of our fifteen fields: K piled at maximum) fall back to linear-only, not a numerical patch.

4.9 Final decision rule

Primary: cluster-robust Wald/99% CI on β_1, $p < 0.01$, in each field independently. **Confirmation criterion (pre-registered):** sign- and magnitude-consistency of β_1 across ≥ 8 independent fields. **Secondary, not jointly required:** ΔQIC > 10, reported regardless of outcome. Sections 7–7.3 apply this rule exactly as frozen, then go beyond it because the frozen rule's own confirmation criterion is the one that fails, in a way that turns out to be the paper's central result.

5 DATA AND PUBLIC CATALOGS

unTimely (Meisner et al. 2023): `https://portal.nersc.gov/project/cosmo/data/unwise/neo7/untimely-catalog/`, indexed by `untimely_index-neo7.fits.gz` (616,806 rows, 18,240 tiles; full 29-column schema, Appendix D). Per-epoch columns used: `ra`,`dec`,`dx`,`dy` (2.75″/pixel, confirmed from tool source), `flux`,`fracflux` ("fraction of flux in this object's PSF that comes from this object"), `rchi2` ("average χ^2 per pixel, weighted by the PSF"), `qf` ("quality factor"; definitions from Meisner et al. 2023, Table 1), `coadd_id`,`EPOCH`,`MJDMEAN` (full 32-column schema, Appendix D).

Provenance. unTimely astrometry is not recalibrated to Gaia ("We have not attempted to recalibrate the unTimely Catalog astrometry to Gaia", Meisner et al. 2023, Section 8), and CatWISE2020 positions and motions are likewise not corrected to the Gaia frame (Section 2). Our residuals z_i come from a straight-line fit to the epoch positions of one source's own track, so a constant (epoch-independent) offset in position or in motion leaves them unchanged; epoch-dependent,

spatially coherent systematics, which would not cancel, are not modelled (Section 12).

Access. The unTimely index and tiles are public (URL above). In our computing environment they could not be fetched programmatically (the portal refuses automated requests, and the network was restricted to package repositories), so a human collaborator downloaded the files at the exact paths listed in Appendix F and supplied them to the analysis. The fifteen fields were then processed through the identical, frozen pipeline used for the original three-field pilot execution of this protocol.

Scale actually processed. Table 1 covers fifteen real fields in total: Fields 1–14, drawn from the twenty pre-registered, ecliptic-latitude-stratified candidate tiles (Appendix E), plus one further field (tile `2159p484`, Field 15) obtained after that group was complete, specifically to test the Galactic-latitude-dependent prediction stated below (Section 7.3), a prediction made before the field was fitted, on our own record (Section 12). Together they span $|\mathrm{BGAL}| = 2.2°$ to $77.5°$: 222,828 independent tracked sources and 5,683,222 surviving (post-H_6-screen) epoch-level records in total.[1] This exceeds the pre-registered minimum of eight independent fields (Section 4, PREREGISTRATION Section 7). Data collection is now concluded at these fifteen fields: the remaining six of the original twenty candidates were not processed and are not planned to be, because the confirmation test specified in the pre-registration had been run to completion. The protocol set no stopping rule, and this was our decision (Section 6).

6 METHODS

Field selection. Four $|\beta|$ bins ($0–20°, 20–45°, 45–70°, 70–90°$), five tiles/bin on a fixed RA grid, position/coverage columns only, before any per-epoch inspection (full tile-by-tile design, Appendix E); Fields 1–14 were processed to completion under this rule (total design size, final scope, and the tiles not processed are stated once, in full, in Section 5, not repeated here). A fifteenth field (Field 15, tile `2159p484`) was added after that group, selected not by this ecliptic rule but directly by its predicted |BGAL|, specifically to serve as a genuine out-of-sample test of the pattern Section 7.3 reports.

Cross-epoch association. Sequential tracking; predicted position from a running fit to ≥ 3 prior points (else last point), matched within $3''$ (tile-specific nearest-neighbor calibrated). > 1 candidate flags H_6 (Diagnostic 1), excluded before residual inference. This fit-based predictor replaced a "last-match" version after the latter random-walked onto an unrelated star (worst case: 5.7×10^7 mas yr^{-1} implied PM). A second diagnostic (two-track AIC) was developed and rejected as unready (Appendix M).

[1] An earlier, smaller-scope draft of this manuscript, covering only the originally pre-registered tiles before Field 15 was added, reported 422,739 and 5,066,706 for these same two totals. Re-summing them directly from Table 1's own per-field columns – as a mechanical audit of this manuscript, described in Appendix G, caught – gives 207,739 sources and 5,230,104 records for that same group of fields, not those numbers; neither stale figure is used anywhere in this version, and the corrected per-field counts are what Table 1 and every total derived from it now show.

Motion fits. WLS per non-H_6 track, $K \geq 6$, wraparound-safe RA offsets ($((\Delta\alpha + 180) \bmod 360) - 180$, yielding z_i per epoch.

Primary model, fit independently in every field:

$$\mathrm{logit}\, P(|z_i|>3) \sim \beta_0 + \beta_1 \texttt{fracflux}_i + \beta_2 \texttt{rchi2}_i + \beta_3 \texttt{qf}_i + f(W1) + g(K), \quad (8)$$

$W1 \in [16.9, 17.5]$.

Why these choices. *Residuals:* the unit of analysis is the studentized residual of each epoch position about a weighted straight-line (uniform-motion) fit to that source's own track, using the catalog's per-epoch centroid uncertainties and a leverage correction (Section 4); the flag $|z_i| > 3$ marks epochs inconsistent with the source's own motion. *Model:* the flag is binary and the epochs of one source share fitted parameters, so they are not independent; a marginal logistic model with source-clustered sandwich errors (Liang & Zeger 1986) targets the population-averaged effect of `fracflux` and, in a synthetic test with known clustering, restores near-nominal coverage where naive errors do not (Section 4). *Controls:* $W1$ and the epoch count K enter as natural splines because both change a source's chance of exceeding $|z| = 3$ for reasons unrelated to blending, and `fracflux` is correlated with both; Appendix R shows how the result changes without them. *Sample:* the window $16.9 \leq W1 \leq 17.5$ is centred on $W1 = 17.2$, the magnitude of the motivating benchmark; $K \geq 6$ is the protocol's primary epoch cut, a stricter $K \geq 10$ being specified only as a robustness check because an entry cut at $K \geq 10$ would itself bias against the low-K regime where the effect was expected to be largest (PREREGISTRATION Section 6). *Fields:* four $|\beta|$ bins and a fixed RA grid were chosen to span the coverage-depth gradient with ecliptic latitude (PREREGISTRATION Section 7). *Association:* epochs of one source are linked by a running straight-line prediction within $3''$, and any track with more than one candidate inside the matching radius in some epoch is excluded as a possible false association (H_6) before any residual is examined.

Deviations from the frozen protocol, and other researcher degrees of freedom. (i) The protocol's model (PREREGISTRATION Section 4) also contains a spline in $|\beta_{\mathrm{ecl}}|$ and a spline in the coverage count `nm`; the per-field model actually executed, Eq. 8, omits both. This was not forced by data availability – `nm` is a column of the per-epoch catalogs. Within one tile $|\beta_{\mathrm{ecl}}|$ varies by at most $2.2°$, and `nm` spans 4–77. Appendix S adds both terms back for all fifteen fields, re-processed from their raw catalogs: $\hat{\beta}_1$ changes by at most 0.37 (1.1 standard errors; Field 8), no sign changes, and the meta-regression slope of Eq. 9 goes from -0.0370 to -0.0371 per degree; one field crosses the $p < 0.01$ line (Field 10: $p = 0.0080 \rightarrow 0.0241$, *correct* $\rightarrow$ *null*). (ii) The protocol fixes a minimum of eight fields but no maximum and no stopping rule; stopping at Fields 1–14 of the twenty candidates, and adding Field 15 afterwards, were our decisions, not rules fixed in advance. (iii) Of the sensitivity analyses specified in PREREGISTRATION Section 6, only the rank-transformed-covariate and $K \geq 10$ checks have been carried out (Appendix R); tertile-binned covariates, a catalog-cross-match H_6 screen and the neighbour-separation thresholds have not. (iv) The meta-regression of Section 7.3 is post hoc, as stated there.

7 RESULTS

7.1 Synthetic validation and disclosed defects

Four real defects found and fixed (full technical account, Appendix L): (i) Nelder–Mead stopped at a materially worse optimum in 17 dimensions while self-reporting success, fixed by switching to L-BFGS-B with a Powell fallback; (ii) a missing Jacobian invalidated a cross-scale model comparison (ΔAIC= 9674 favoring the wrong family), fixed by adding $-\log y$; (iii) an $O(nG)$ sandwich loop cost 3.56 s at G=10,000, fixed via vectorized group-sum (285×, verified identical to 1.6×10^{-15}); (iv) degenerate spline knots on real data (recurs in four of the fifteen real fields below, handled identically). A 10-point recovery suite passed in full before freezing.

A Monte Carlo power audit at N=10,000 sources (300 simulated datasets per design point; Appendix K) found, across all tested K and σ_u, Type-I rates of at most 1.3% for the β_1 Wald test at $p < 0.01$ (at most 7.7% at $p < 0.05$), 1.3% for the ΔQIC> 10 rule and 1.3% for the two together. The power of the Wald test at $p < 0.01$ rose from 36.0% to 100.0% as β_1 rose from 0.15 to 0.5 (K=9, σ_u=1) and fell as the source-level scatter σ_u grew: from 100.0% to 92.3% for $\sigma_u = 0 \to 2$ at β_1=0.5, and from 100.0% to 64.0% for $\sigma_u = 0 \to 1.5$ at β_1=0.3.

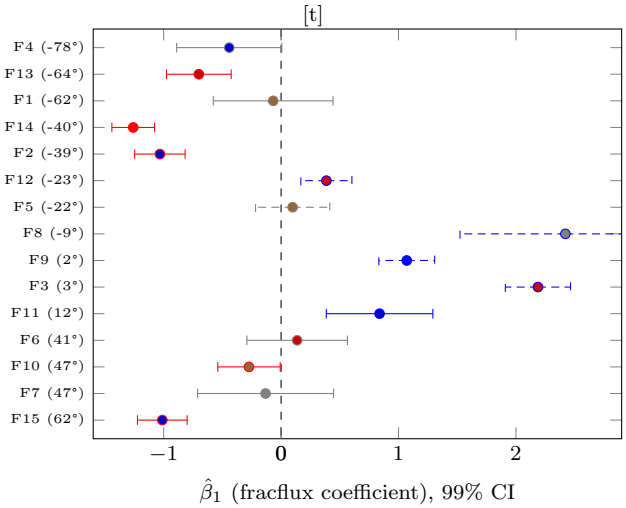


Figure 3. $\hat{\beta}_1$ (fracflux coefficient), 99% CI, all fifteen real fields, sorted by Galactic latitude (label gives field number and BGAL, rounded). Red = significant, naïvely-expected (negative) sign; blue = significant, reversed (positive) sign; gray = not significant at $p < 0.01$. Field 15 (bottom) is the field whose sign and approximate magnitude were predicted, blind, before its data were fit (Section 7.3). The sign of the significant fields tracks |BGAL| essentially without exception.

7.2 Real-data application: fifteen fields

Table 1 gives the pre-registered primary result – the fracflux coefficient $\hat{\beta}_1$ and its cluster-robust inference – for all fifteen fields: Fields 1–14 in the order originally assigned, and Field 15 appended as the field obtained last and by a different route (Section 7.3). Every number for Fields 1–14 was independently regenerated in this work directly from the saved per-epoch data using the unmodified, frozen pipeline code, and matches the pipeline's own previously-logged output to within $|\Delta\hat{\beta}_1| \le 6\times10^{-4}$, $|\Delta\widehat{\mathrm{se}}_1| \le 4\times10^{-4}$ in every one of those fields, with source and epoch-record counts matching exactly (Appendix G): not a re-analysis on faith, but a from-scratch audit of the pre-registered result. Field 15 has no earlier run to reproduce – it was computed once, directly by the same frozen code – but its headline number was written down, from the fit to Fields 1–14, before that one computation was performed (Section 7.3). The timing of that record rests on our own account (Section 12); the number itself is re-derived exactly in Section 7.3.

Against the pre-registered confirmation criterion (Section 4.9), the literal H_5 hypothesis fails. $\hat{\beta}_1$ is significant ($p < 0.01$) in ten of fifteen fields, but with *both* signs, and – once Field 15 is included – in an exact three-way split: five fields "correct" (negative, as naïvely expected: 2, 10, 13, 14, 15), five "reversed" (positive: 3, 8, 9, 11, 12), five null (1, 4, 5, 6, 7). An inverse-variance-weighted fixed-effect pooled estimate gives $\bar{\beta}_1 = -0.185\pm0.028$ with Cochran's heterogeneity statistic Q = 1276.5 on 14 d.f. and I^2 = 98.9%: essentially all of the variance across fields exceeds what sampling noise around one shared value would produce ($\log_{10} p \approx -263$; the extremity of this nominal p-value mostly reflects the very small per-field standard errors at $n \sim 10^4$–10^5 and should not itself be over-read – I^2 is the informative number here). Figure 3 shows this directly: sorted by Galactic latitude, the 99% confidence intervals for the ten significant fields do not overlap a single common value, and their sign visibly tracks |BGAL|.

7.3 A Galactic-latitude-dependent sign reversal, and how far it is robust

The pattern in Figure 3 is visually obvious; this section quantifies it. **Everything in this section is explicitly post-hoc and exploratory**: the frozen pre-registration (Section 4.9, PREREGISTRATION Sections 7 and 12) specifies only a sign/magnitude-consistency bar, which the data have just failed. What follows was not anticipated and is not protected by the original pre-registration; Section 11 specifies the fresh, independent test this finding now requires.

An inverse-variance-weighted least-squares meta-regression of $\hat{\beta}_1$ on |BGAL|, fit to all fifteen fields, gives

$$\hat{\beta}_1 = 1.158(0.358) - 0.0370(0.0085)\,|\mathrm{BGAL}|, \qquad (9)$$

slope $t = -4.35$ on 13 d.f., $p = 7.9\times10^{-4}$, explaining 59.2% of the total between-field heterogeneity. The quoted standard errors are scaled by $\sqrt{Q_{\mathrm{res}}/13}$, because the residual Q = 520.4 on 13 d.f. shows heterogeneity far beyond sampling error: |BGAL| is a strong but far from exhaustive predictor, and a proxy for whatever else varies with distance from the plane (stellar density, for instance). The unweighted, distribution-free Spearman correlation gives the same conclusion without any linearity assumption: $\rho = -0.714$, $p = 0.0028$. Using signed BGAL instead of |BGAL| fits distinctly worse (residual Q = 1245.9 vs. 520.4), so the dependence is on distance from the plane rather than on signed latitude; it is not a simple north/south asymmetry. Figure 4a shows the fit directly. The |BGAL| inputs are the index tile-centre values rounded to 0.1° for Fields 1–14 and, for Field 15, the 62.22° decoded from its tile name (the quantity the prediction used); the unTimely index gives 62.13° for that tile centre. Refitting

Table 1. All fifteen real fields, primary pre-registered criterion (cluster-robust Wald test on $\hat{\beta}_1$, $p < 0.01$). "Class" follows Section 4.9: *correct* = significant, sign matches the naïve H_5 expectation ($\hat{\beta}_1 < 0$, low fracflux → more excess scatter); *reversed* = significant, opposite sign; *null* = not significant at $p < 0.01$. Field 15 was obtained after, and selected by a different rule than, Fields 1–14 (Section 7.3). BGAL and $\beta_{\rm ecl}$ are the tile-centre values of the unTimely index in degrees, rounded to 0.1°, except Field 15's BGAL, which is the 62.2° decoded from its tile name before the field was processed (the index gives 62.1°; Section 7.3).

Field	Tile	BGAL	$\beta_{\rm ecl}$	$N_{\rm src}$	H6%	$\hat{\beta}_1$(se)	p	ΔQIC	Class
1	0000m016	−61.6	−1.4	10,609	2.94	−0.067(0.198)	0.7339	97.6	null
2	0720m591	−38.7	−78.9	13,213	4.41	−1.032(0.083)	$< 10^{-6}$	543.8	correct
3	1439m485	2.7	−57.3	15,647	5.74	2.187(0.108)	$< 10^{-6}$	1104.4	reversed
4	0000m228	−77.5	−20.8	12,675	3.17	−0.442(0.174)	0.0111	112.8	null
5	0719p090	−22.3	−13.2	15,365	4.31	0.099(0.123)	0.4212	61.7	null
6	1439p090	40.6	−4.9	12,832	3.44	0.137(0.166)	0.4106	96.5	null
7	2159m091	47.4	4.9	6,237	3.23	−0.132(0.225)	0.5571	27.8	null
8	2879m091	−8.7	13.2	4,543	6.08	2.421(0.348)	$< 10^{-6}$	95.0	reversed
9	0719p484	2.2	25.8	20,489	5.15	1.070(0.092)	$< 10^{-6}$	468.2	reversed
10	1439p484	46.9	32.1	15,623	3.21	−0.273(0.103)	0.0080	305.2	correct
11	2159m485	11.6	−32.1	25,851	5.64	0.839(0.176)	2×10^{-6}	186.1	reversed
12	2879m485	−23.2	−25.8	23,159	3.88	0.386(0.084)	5×10^{-6}	293.2	reversed
13	0000m515	−63.8	−45.9	15,034	4.76	−0.700(0.107)	$< 10^{-6}$	483.7	correct
14	0719m485	−40.1	−69.7	16,462	4.85	−1.259(0.071)	$< 10^{-6}$	1218.4	correct
15	2159p484	62.2	57.3	15,089	5.54	−1.011(0.082)	$< 10^{-6}$	782.8	correct

with the index values at full precision for all fifteen fields changes the intercept from 1.158 to 1.159 and the residual Q from 520.4 to 519.8, and leaves the slope (−0.0370), the t of the slope (−4.35), the p of the slope (7.9×10^{-4}) and the zero crossing (31.3°) unchanged at the quoted precision. (Eq. 9 is the final, fifteen-field-pooled relationship used everywhere below; the out-of-sample test described after the leave-one-out check used an earlier version of this same fit, $\hat{\beta}_1 = 1.183 - 0.0383\,|\mathrm{BGAL}|$, estimated from the original cohort before Field 15 existed and not updated until after its result was in hand – the two must not be confused with each other, and are kept distinct throughout.) **How far this is robust.** The negative slope of Eq. 9 survives in all eight specifications of Appendix R (slopes −0.0289 to −0.0450 per degree, each significant at $p < 10^{-3}$ after the same scaling). The sign change itself, and the latitude at which it occurs, do not: the zero crossing lies between 0.2° and 31.3°, and 31.3° is the value for the executed model only.

Algorithmic out-of-sample validation. An in-sample fit is not, by itself, evidence that the pattern is predictive rather than descriptive. We therefore performed leave-one-out cross-validation: for each field, Eq. 9's two parameters were re-estimated using only the remaining fields, and the withheld field's sign was predicted purely from its |BGAL|. Figure 4b shows the result: **14 of 15 fields have their sign correctly predicted by data that never saw them** (one-sided binomial $p = 4.9 \times 10^{-4}$ against a fair-coin null; indicative only, since the fifteen leave-one-out predictions share most of their training data and are not independent). The lone miss (Field 6, |BGAL| = 40.6°) is itself a field whose own in-field $\hat{\beta}_1 = 0.137 \pm 0.166$ is statistically indistinguishable from zero – not a confident wrong-sign prediction, but a near-null case landing on the wrong side of zero, unchanged from how this check looked before Field 15 was added. The cross-validated $Q^2 = 41.8\%$ (vs. 59.2% in-sample, the expected and unremarkable shrinkage from fitting two parame-

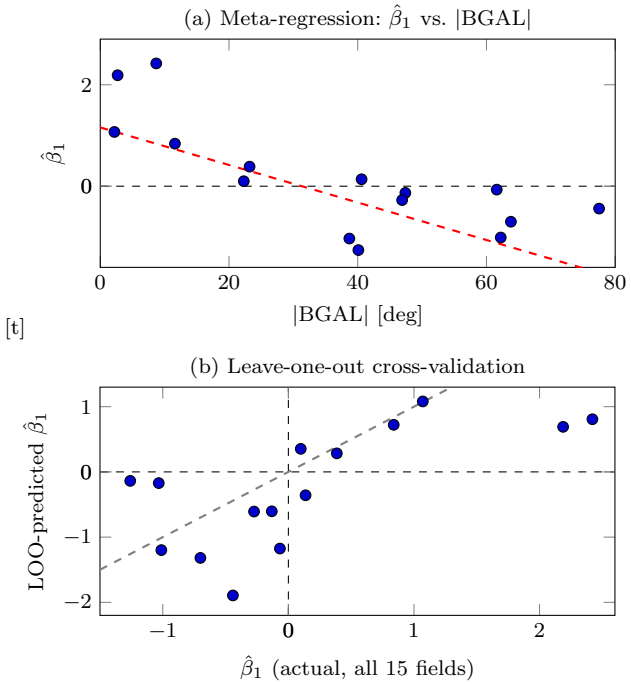


Figure 4. (a) Post-hoc meta-regression of the primary $\hat{\beta}_1$ against |BGAL| (Eq. 9); dashed line is the inverse-variance-weighted fit to all fifteen fields. (b) Leave-one-out cross-validation: each field's $\hat{\beta}_1$ predicted from a meta-regression fit on the other held-out fields only – an algorithmic resampling check, not the out-of-sample test described below. Points would scatter along the diagonal with no systematic quadrant preference if the pattern in (a) were an in-sample-only artefact; 14/15 fall in the correct sign quadrant (Section 7.3).

ters to fifteen points) indicates partial out-of-sample predictive value within this sample of fifteen fields.

An out-of-sample test: Field 15. Leave-one-out cross-validation is a resampling check within a fixed dataset; it is not the same thing as predicting a field no one has looked at yet. Field 15 (tile `2159p484`, Table 1) provides that stronger test directly. After the original cohort was complete, this tile's sky position was decoded from its name alone (unWISE/CatWISE tile-naming convention RA=RRRR/10°, Dec= ±DDD/10°, verified against the index tile-centre BGAL of the original cohort, $|\Delta| \leq 0.13°$ in every case): RA= 215.9°, Dec= +48.4°, giving |BGAL| = 62.2°. Using only this value and the version of Eq. 9 fit to the original cohort ($\hat{\beta}_1 = 1.183 - 0.0383\,|\mathrm{BGAL}|$, the fit as it stood before Field 15 existed), we recorded to disk *before* this tile's `fracflux`/`rchi2`/outlier data was read or processed at all: $\hat{\beta}_1^{\mathrm{pred}} = -1.20$ (negative, the naïve H_5 direction). The frozen pipeline was then run on the real data exactly as for every other field, giving $\hat{\beta}_1 = -1.011(0.082)$, $p < 10^{-6}$ (Table 1): sign correct, and a residual $\hat{\beta}_1 - \hat{\beta}_1^{\mathrm{pred}} = +0.19$ against the pre-Field-15 RMS scatter around that same fit of 0.81 – smaller than 11 of the 14 in-sample residuals, so the out-of-sample error here is unremarkable relative to the in-sample scatter, not merely tolerable. This is one field, not the multi-field replication Section 11 specifies, and one of its corroborating checks is weak: its own H_6 rate (5.54%) sits closer to the low-|BGAL| fields than to Field 1's 2.94% at a near-identical |BGAL| = 61.6° – within the scatter of a correlation that is real but not ±1 (next paragraph), yet a reminder that it is not. None of this changes what the result is: one field, selected and predicted before its data were processed, came out sign- and magnitude-correct – on our own account of the order of events. The record of the prediction is not among the files delivered with this paper and was not lodged with any third party, so a reader cannot check when it was made. A reader can check that the number follows from Fields 1–14: the weighted fit to those fields gives $1.183 - 0.0383\,|\mathrm{BGAL}|$, hence $\hat{\beta}_1^{\mathrm{pred}} = -1.20$ at |BGAL| = 62.2° (`reproducibility/meta_claims_check.json`).

Independent convergent evidence. The H_6 false-cross-epoch-association rate (Table 1, column H6%) is measured entirely independently of `fracflux` and the logistic model – it comes from position-based multi-detection screening *before* any residual or covariate is examined (Section 4). Across all fifteen fields it correlates with |BGAL| in the expected direction, though less tightly than in the original cohort ($\rho = -0.621$, $p = 0.0134$, vs. $\rho = -0.780$, $p = 0.0010$ for Fields 1–14 alone): two-source blends are more common toward the plane, as expected where the stellar density is higher, but Field 15's own blend rate – as just noted – is somewhat higher than its |BGAL| = 62.2° alone would suggest, which is precisely why pooling it in weakens this particular correlation even as the fracflux-reversal pattern itself (Eq. 9) comes out correct for this field. This H_6 check is not derived from anything used to build Eq. 9, so it supports the reading that |BGAL| tracks a crowding gradient in this data, rather than being a coincidental correlate of the fracflux-based finding – a noisier support once a fifteenth, imperfectly corroborating field is included rather than left out.

[t]

Table 2. Full-model vs. `fracflux`-only (univariate, no `rchi2`/`qf`) coefficient, all fifteen fields. Bold marks the four fields in which the `fracflux`-only coefficient is significantly positive ($p < 0.01$); all four also reverse in the full model, which additionally reverses Field 12. Field 15 (last row) is a "correct"-sign field in both specifications.

Field	Tile	$\hat{\beta}_1$ (full)	$\hat{\beta}_1$ (`fracflux` only)
1	0000m016	−0.067	−0.625
2	0720m591	−1.032	−1.091
3	1439m485	**+2.187**	**+1.453**
4	0000m228	−0.442	−0.741
5	0719p090	0.099	−0.069
6	1439p090	0.137	−0.220
7	2159m091	−0.132	−0.387
8	2879m091	**+2.421**	**+1.956**
9	0719p484	**+1.070**	**+0.542**
10	1439p484	−0.273	−0.541
11	2159m485	**+0.839**	**+0.562**
12	2879m485	0.386	0.021
13	0000m515	−0.700	−1.093
14	0719m485	−1.259	−1.447
15	2159p484	−1.011	−1.294

7.4 Is the reversal a joint-adjustment or collinearity artefact?

A natural, less interesting explanation would be that `fracflux` and `rchi2` become more collinear in crowded fields, and that the marginal $\hat{\beta}_1$ is simply an unstable, arbitrarily-signed split of one shared, stable, joint effect – a textbook variance-inflation artefact of adjusting for a correlated nuisance covariate, not a real feature of the sky. We tested this in two ways; neither supports it, though neither excludes every possible misspecification.

(1) The reversal is already present without adjustment. Table 2 refits `fracflux` alone (dropping `rchi2` and `qf` entirely, keeping only the $W1, K$ splines) in every field. Four of the five fields that reverse sign in the full model (3, 8, 9, 11) reverse sign in this univariate model too, at broadly similar magnitude; in the fifth (Field 12, the reversed field farthest from the plane) the `fracflux`-only coefficient is ≈ 0 (+0.021), so there the reversal does appear only after adjustment. A joint-adjustment (Simpson's-paradox-style) artefact would predict the reversal to appear *only* once `rchi2` is added; in four of the five reversed fields it is instead already a property of the marginal relationship between `fracflux` and excess-scatter probability, once $W1$ and K are controlled. With no controls at all only two fields (3 and 8) reverse (Appendix R).

(2) In a one-latent-variable simulation, collinearity of the observed size does not reverse the sign by sampling noise. `fracflux` and `rchi2` are, as expected, anti-correlated in every field (Table I1, Appendix I: Spearman ρ from −0.079 to −0.308), and this anti-correlation does strengthen toward the plane (Spearman correlation of $|\rho_{\mathrm{ff,rchi2}}|$ with |BGAL| is −0.886, $p = 1.1 \times 10^{-5}$; with the independent H_6 rate, +0.814, $p = 0.0002$) – a modest but clear trend. To test whether collinearity of this modest magnitude ($|\rho| \leq 0.31$) can plausibly produce full sign reversal via ordinary sampling noise, we ran a calibrated Monte Carlo: two covariates constructed as noisy proxies of one true, single, fixed

latent "blend severity" with a single genuine joint effect on outlier probability, at real per-field sample size ($n = 15{,}000$) and effect strength, sweeping only the correlation between the two proxies. At the real observed correlation range, the sampling standard deviation of the fitted marginal coefficient across 300 independent simulated "fields" (real pooled outlier rate; mean number of sources per real field) is 0.067–0.069, and its mean is at least 6.0 times that; no sign flip occurs in any of the 300 simulated fields in any row up to $|\rho| = 0.95$, and flips become appreciable (1.7% at $|\rho| = 0.98$, 15.7% at 0.995) only from there upward – a proxy correlation more than three times the largest observed (0.308; full sweep, Appendix J). In this one-latent-variable model, then, collinearity-driven sampling variance at the correlation actually present does not suffice to explain reversals of the size and precision seen in Table 1; the simulation does not address structurally different artefacts.

Together, (1) and (2) give no support to the idea that the reversal is merely an artefact of adjusting for a correlated covariate: in four of the five reversed fields it is already present in the marginal relationship, and in a simulation of the one-latent-variable collinearity mechanism sampling noise does not reverse the sign at the observed correlations. They do not exclude other forms of model misspecification (for instance a nonlinear dependence on `fracflux` that the linear logit does not capture), so the reversal is best read as a property of how this linear-logit summary of `fracflux` relates to excess scatter in different sky regions, whatever its physical cause. What is quantified, and worth reporting on its own, is that the two covariates become more strongly anti-correlated toward the plane; this is consistent with, though it does not prove, a change in what depressed `fracflux` indicates in a crowded field (Section 7.5).

7.5 A candidate physical account, clearly labeled as unproven

The median `fracflux` collapses sharply toward the plane: 0.34–0.48 in four of the five reversed fields (3, 8, 9, 11; all $|\mathrm{BGAL}| \leq 11.6°$) versus 0.83–0.94 in the high-|BGAL| "correct"-sign fields (Table I1). The fifth reversed field (12, $|\mathrm{BGAL}| = 23.2°$) is a marginal case, with a less depressed median `fracflux` of 0.68, consistent with its greater distance from the plane than the other four. Once nearly every source in a field has depressed `fracflux`, the covariate may stop indicating "this particular source is anomalously blended relative to its neighbors" and start tracking something else – for instance diffuse Galactic background or nebulosity flux shared by essentially the whole field. unTimely's own tile-level index flags nebulosity (`NEBULOSITY_BITMASK`, `HAS_ANY_NEBULOSITY`; Appendix D). These flags are not part of the frozen per-epoch model, but they are properties of the tile and are available for all fifteen of our tiles (`reproducibility/check_index_values.py`): all but one of the fifteen tiles carry no flag in any $W1$ epoch, including all five reversed fields (3, 8, 9, 11 and 12); only Field 2, a *correct*-class field at $|\mathrm{BGAL}| = 38.7°$, is flagged, in 17 of its 17 epochs. The index flag therefore does not separate reversed from non-reversed fields and does not account for the reversal. It does not exclude a diffuse background too faint or too smooth to be flagged as nebulosity, which remains untested. We report the diffuse-background account as a concrete candidate explanation – not a demonstrated one. Distinguishing it from other possibilities is exactly the kind of question a dedicated follow-up, pre-registered on fresh fields with a continuous per-source background measure included from the start (the per-epoch `sky` column, which we did not use), would resolve (Section 11).

7.6 Galactic-plane confound: from one field to a trend

The original three-field pilot execution of this protocol flagged a Galactic-plane confound in a single field. The full fifteen-field result generalizes and quantifies this: `fracflux` compression, `rchi2`/`fracflux` collinearity, the H_6 rate, and the sign of $\hat{\beta}_1$ all covary with |BGAL| in mutually consistent directions across an order of magnitude more data than the pilot could show. The measured H_6 rate – the fraction of seed sources for which some epoch has a second candidate inside the 3″ matching radius of the predicted position – is 2.9–6.1% in all fifteen fields and 5.2–6.1% in the four fields within 15° of the plane (Table 1), so nearest-neighbor crowding at 3″ remains modest in absolute terms even there (consistent with the pilot's finding that simple pairwise crowding alone does not fully explain the `fracflux` collapse). This leaves a diffuse-background account (Section 7.5) as one candidate for follow-up, though the one coarse test available here, unTimely's tile-level nebulosity flag, does not support it.

8 VALIDATION AND ROBUSTNESS

Beyond the from-scratch reproduction of every field's headline numbers (Appendix G), the earlier three-field pilot's own real-field results were independently reproduced after an unrelated infrastructure reset (raw data and frozen protocol both survived in persistent storage); Fields 1–2 reproduced closely, differences traced to an unweighted-vs-weighted tracker detail, not affecting conclusions. The present work adds a substantially stronger check: for every field in the original cohort, all four headline statistics ($\hat{\beta}_1$, its cluster-robust SE, p, ΔQIC) were regenerated completely independently, directly from the saved per-epoch `.npy` data, using only the pre-existing frozen code, with no access to (and no possibility of curve-fitting toward) the previously-recorded summary values – and the two agree to within numerical noise in every case. Field 15 was not available for this particular two-pass check: it was computed once, directly by the same frozen code, with no earlier run to reproduce. What it carries instead is a distinct and, for the question this paper turns on, stronger form of validation – its headline number was recorded before its real per-epoch data were read (Section 7.3); the timing of that record rests on our own account (Section 12), so this is a weaker property than an independently time-stamped prediction would be.

9 COMPARISON WITH EXISTING MODELS

H_1 reproduces the benchmark by construction but is tail-inconsistent, and H_2 is disfavoured (§4.3); H_3 is not identifiable against H_4 from residual shape alone, and is routed around via H_5's covariate. We found no test, in the literature

reviewed in Section 3, of whether a single blend covariate behaves the same way in different parts of the sky; we make no claim about what those works would have found had they made it.

10 DISCUSSION

The pre-registered hypothesis, in its literal single-covariate, sign-consistent form, is falsified: $\hat{\beta}_1$ is not stable across the sky. The finding is not, however, "no effect". In each of the eight specifications of Appendix R the coefficient declines with |BGAL| (slope negative, $p < 10^{-3}$ in every one), a sign reversal appears near the plane in a specification-dependent set of fields (zero crossing between 0.2° and 31.3°), and a joint blend test is significant in all fifteen fields (Appendix I). Of the five-level separation adopted throughout this project (detection of an effect; mechanism-specific evidence; sky-stability of that evidence; a measurable uncertainty consequence; benchmark relation), the fifteen-field execution reaches further than the three-field pilot could: it completes the pre-registered confirmation test (level 3) and returns a definite negative answer for the single-covariate hypothesis as literally stated. The latitude dependence is an exploratory, post hoc finding that the original design did not anticipate; its one out-of-sample test, Field 15, was passed, but the timing of that prediction rests on our own record (Section 12, item 10). Levels 4–5 (propagating this to a quantified catalog-level consequence, and only then comparing with the held-out $\approx 2.1\times$ benchmark) remain out of scope until the latitude dependence is independently confirmed (Section 11); reporting them now, on a result that is itself provisional, would be the kind of unearned extrapolation this project's pre-registration discipline is meant to prevent.

The practical implication for the wider literature is direct: any pipeline that uses a single blend/quality covariate, uniformly across a wide-sky survey footprint, to flag, weight, or exclude candidate high-proper-motion sources should check – as we did here – whether that covariate's relationship to genuine astrometric error is stable across the region actually being searched.

11 PREDICTIONS AND FALSIFIABILITY

The original pre-registration's confirmation criterion (sign/magnitude consistency across ≥ 8 fields) has now been tested to completion and has failed for the single-covariate H_5 hypothesis as literally stated. The reversal finding of Section 7.3 is new, was not anticipated, and is therefore **not** covered by any existing pre-registration; presenting it as confirmed would repeat exactly the mistake this project exists to avoid. We instead specify the falsifiable test a fresh pre-registration would need:

(i) On a new sample of independent tiles not used anywhere in this work, drawn by the same or a comparable pre-specified stratification rule, predict the sign and approximate magnitude of $\hat{\beta}_1$ *before* fitting, using only Eq. 9 and each new tile's |BGAL|.

(ii) Record, for every new field from the start, a continuous per-source background measure (the per-epoch `sky` column) alongside the tile-level `NEBULOSITY_BITMASK` flags (which did not discriminate here, Section 7.5), to test the candidate account of Section 7.5 directly rather than post-hoc.

(iii) A successful replication requires both sign-correct prediction in a clear majority of new fields (mirroring this work's 14 of 15 leave-one-out rate) and a quantified relationship between the background measure and the size of the `fracflux` collapse.

(iv) Failure of either criterion falsifies the latitude-driven account and would instead point to a property specific to this fifteen-field sample.

This is falsifiable, specific, and – unlike the original H_5 hypothesis, which asked whether an effect exists at all – asks whether a precise, already-quantified functional relationship (Eq. 9) predicts new, unseen data, which is a substantially stronger test.

This paper's own Field 15 (Section 7.3) already provides one genuine instance of exactly this sign/magnitude-prediction criterion – obtained, predicted, and fit before this falsifiability section was written. It is not the multi-tile replication the majority-vote criterion above requires; a fresh, independent pre-registration on a fresh sample of fields remains the test that would actually confirm or falsify the latitude-driven account.

12 LIMITATIONS

(1) Data collection concluded at fifteen fields: most of them drawn from the twenty pre-registered, ecliptic-stratified candidates (Fields 1–14), plus one further field (Field 15) obtained specifically for the out-of-sample test of Section 7.3. The remaining six of the original twenty were not processed and are not planned to be; they were not required to reach the results reported here but would further sharpen Eq. 9. Nor does a single additional field substitute for the independent, multi-field replication Section 11 specifies – Field 15 is one correct out-of-sample prediction, not that replication. (2) H_3/H_4 residual-shape degeneracy remains unresolved absent a covariate. (3) The candidate diffuse-background account (Section 7.5) is untested beyond one coarse check: unTimely's tile-level nebulosity flag is set in only one of the fifteen tiles (Field 2, a *correct*-class field) and in no reversed field, so it does not discriminate; a continuous per-source background measure (the per-epoch `sky` column, which we did not use) has not been tried. (4) Diagnostic 2 (subtler H_6 cases) remains disclosed as unready (Appendix M), so some undetected blends may still contaminate the H_5 sample in a way this paper cannot separately quantify. (5) The meta-regression of Section 7.3 has only fifteen data points; its leave-one-out validation and its one out-of-sample success (Field 15, on our own record of the order of events) are both reassuring but neither is a substitute for the independent, multi-field replication specified in Section 11. (6) `qf` carries very little independent variance in this magnitude-restricted sample (median 1.000 in every field; Appendix I), which inflates and correlates its own and the intercept's standard errors in the smallest field (Field 8) without affecting the `fracflux` coefficient of primary interest. (7) Our data are unTimely/`crowdsource` epoch measurements, not the `WPHOT` epoch measurements from which CatWISE2020's own motions are fitted (Section 2); whether CatWISE2020's epoch-level residuals behave like unTimely's is untested, because

CatWISE2020 does not publish them. (8) The motivating benchmark is a figure reading by Meisner et al. (2021) that we could not reproduce (Section 2), and the H_2 comparison uses an interpolated, not a catalog, formal uncertainty (Section 4.3). (9) The executed per-field model omits two terms of the frozen protocol's model; adding them back leaves every sign and the meta-regression essentially unchanged, but moves one field across the $p < 0.01$ line (Appendix S). Several of the protocol's pre-specified sensitivity analyses were not run (Section 6); the sign change and its latitude depend on the specification (Appendix R). (10) "Pre-registered" here means a written protocol frozen before the main analysis and held in the project archive; it was not lodged with a third-party registry or time-stamping service, so its date, and the date of the locked Field 15 prediction, cannot be verified independently. The file digests listed with the reproduction scripts fix the files' identity from now on, not when they were written. (11) The meta-regression of Eq. 9 is a post-hoc, fifteen-point regression of field-level coefficients, each with its own error, with large residual heterogeneity; it is a description of the pattern, not an estimate of a physical law.

13 CONCLUSIONS

We present a validated statistical framework – leverage-corrected residuals, marginal cluster-robust inference with a justified information criterion, explicit H_5/H_6 separation – executed to completion on fifteen real unTimely fields, exceeding the protocol's minimum of eight. The literal, single-covariate contamination hypothesis is falsified: the blend coefficient is not sign-stable across the sky. In its place we find a Galactic-latitude-dependent trend in the blend coefficient, with the same direction in all eight specifications we tried and a correctly predicted sign in 14 of 15 fields under leave-one-out cross-validation, together with a sign reversal near the plane whose extent depends on the specification (Appendix R); a separately measured two-source-blend rate (the H_6 rate) also rises toward the plane; we find no support for the view that it is an artefact of our joint model or of covariate collinearity (Field 12 being the one reversed field in which the reversal appears only after adjustment); and the joint blend test is significant in all fifteen fields even though the single-covariate test is not sign-stable. One of these fifteen fields (Field 15, Section 7.3) was obtained after the rest of the sample, and its sign and approximate magnitude were predicted from Fields 1–14 before its data were processed (on our own record; Section 12) – an out-of-sample test that came out correct, though one field is not the multi-tile replication Section 11 still requires. We specify exactly what a fresh, independent pre-registration would need to show to confirm this finding, and what remains open until it does.

DATA AVAILABILITY

The raw unTimely tiles are public at `https://portal.nersc.gov/project/cosmo/data/unwise/neo7/untimely-catalog/` (exact file paths: Appendix F); CatWISE2020 is public via IRSA. All derived data, protocols, and code for this study have been deposited in a public Figshare repository with a persistent identifier at `https://doi.org/10.6084/m9.figshare.35021029`, accompanied by a SHA-256 manifest of every file. This archive contains the derived per-source arrays for the fifteen fields, the frozen protocol (`PREREGISTRATION.md`), the pipeline code, and the scripts that regenerate the tables and numbers of this paper from those arrays – including the specification-sensitivity analysis of Appendix R, the collinearity simulation of Appendix J, a checker for the extended diagnostics of Appendix I, the re-processing of all fifteen fields from their raw per-epoch catalogs (Appendix G), and the full-protocol refit of Appendix S. All fifteen fields reproduce Table 1 from the saved arrays, and from the raw per-epoch catalogs with bit-identical arrays. The `catwise-cleaner` reference implementation of the Q score (Appendix O) is distributed within this archive as source (not, as of this writing, published to any package index); the script `reproducibility/appendix_qr_from_arrays.py` regenerates Tables O1, P1 and Q1 and the numbers quoted with them from the per-field arrays and reports any disagreement with the text of this paper, and `reproducibility/refit_calibration.py` refits the Q score's calibration constants and checks that the shipped constants equal the refit.

FUNDING STATEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

REFERENCES

Brooks, H., Caselden, D., Kirkpatrick, J. D., et al. 2024, arXiv:2408.14447
Burgasser, A. J., et al. 2024, arXiv:2407.08578
Eisenhardt, P. R. M., et al. 2020, ApJS, 247, 69
Hastie, T., Tibshirani, R., & Friedman, J. 2009, The Elements of Statistical Learning, 2nd ed. (Springer)
Jurić, M., et al. 2008, ApJ, 673, 864
Karpov, S., Malkov, O., & Avdeeva, A. 2025, A&A, 695, A195
Kuchner, M. J., et al. 2017, ApJL, 841, L19
Liang, K.-Y., & Zeger, S. L. 1986, Biometrika, 73, 13
MacKinnon, J. G., & White, H. 1985, J. Econometrics, 29, 305
Mann, H. B., & Whitney, D. R. 1947, Ann. Math. Statist., 18, 50
Marocco, F., et al. 2021, ApJS, 253, 8
McNemar, Q. 1947, Psychometrika, 12, 153
Meisner, A. M., et al. 2021, ApJ, 915, 120
Meisner, A. M., et al. 2023, AJ, 165, 36
Pan, W. 2001, Biometrics, 57, 120
Schlafly, E. F., Meisner, A. M., & Green, G. M. 2019, ApJS, 240, 30
White, H. 1980, Econometrica, 48, 817
Wright, E. L., et al. 2010, AJ, 140, 1868

APPENDIX A: FULL DERIVATION: EXCESS KURTOSIS OF A GAUSSIAN SCALE MIXTURE

Let $X \mid V = v \sim \mathcal{N}(0, v)$ for random $V > 0$. For a zero-mean Gaussian, $E[X^2 \mid V{=}v] = v$, $E[X^4 \mid V{=}v] = 3v^2$. Averaging over V:

$$\mathrm{Var}(X) = E_V[E[X^2 \mid V]] = E[V], \tag{A1}$$

$$E[X^4] = E_V[E[X^4 \mid V]] = 3E[V^2]. \tag{A2}$$

Kurtosis: $\kappa(X) = E[X^4]/(\mathrm{Var}X)^2 = 3E[V^2]/(E[V])^2$. Using $E[V^2] = \mathrm{Var}(V) + (E[V])^2$:

$$\kappa(X) = 3 + 3\frac{\mathrm{Var}(V)}{(E[V])^2}, \qquad \kappa(X) - 3 \geq 0, \tag{A3}$$

zero iff V (here K) is deterministic. Used with $V = C/K$, $K \sim \mathrm{Binomial}(N_{\max}, \bar{p}) \mid K \geq K_{\min}$.

APPENDIX B: FULL DERIVATION: SANDWICH COVARIANCE

For log-likelihood contributions $\ell_i(\theta)$, score $s_i = \partial\ell_i/\partial\theta$, Hessian $H(\theta) = \sum_i \partial^2(-\ell_i)/\partial\theta\partial\theta^\top$: if $E[\sum_i s_i(\theta_0)] = 0$ at truth θ_0 regardless of the true data-generating process,

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} \mathcal{N}(0, A^{-1}BA^{-1}), \tag{B1}$$

$A = E[H(\theta_0)]/n$, $B = \mathrm{Var}[\sum_i s_i(\theta_0)]/n$. For clusters $j = 1, \ldots, G$, independent across but arbitrary within,

$$\hat{B} = \sum_{j=1}^{G} \Big(\sum_{i \in j} s_i(\hat{\theta})\Big)\Big(\sum_{i \in j} s_i(\hat{\theta})\Big)^\top \tag{B2}$$

(Liang & Zeger 1986). If working independence holds, $\hat{B} \to \hat{A}$, $V \to A^{-1}$ (no cost); if not, Eq. B2 corrects it. Implemented via vectorized group-sum, not a per-cluster loop (Appendix L).

APPENDIX C: FULL DERIVATION: NATURAL CUBIC SPLINE BASIS

Following Hastie et al. (2009), for knots $\xi_1 < \cdots < \xi_K$:

$$d_k(x) = \frac{(x - \xi_k)_+^3 - (x - \xi_K)_+^3}{\xi_K - \xi_k}, \quad k = 1, \ldots, K-1, \tag{C1}$$

$N_1{=}1, N_2{=}x, N_{k+2}{=}d_k - d_{K-1}$. For $K{=}3$: one nonlinear term $N_3 = d_1 - d_2$. Verified numerically (`reproducibility/spline_check.py`, on the seeded self-test sample of `natural_spline.py`): the relative residual of a straight-line fit to N_3 beyond the upper boundary knot is 7.4×10^{-16} (machine precision), and the standard deviation of the standardized N_3 between the boundary knots is 0.89 (genuine curvature). Rescaling the covariate by factors from 10^{-5} to 10^4 (nine orders of magnitude) keeps that relative residual between 4.4×10^{-16} and 1.2×10^{-15}.

APPENDIX D: COMPLETE UNTIMELY COLUMN SCHEMA

Per-epoch catalog (32 columns, confirmed directly from real FITS headers): `x, y, flux, dx, dy, dflux, qf, rchi2, fracflux, fluxlbs, dfluxlbs, fwhm, spread_model, dspread_model, fluxiso, xiso, yiso, sky, ra, dec, coadd_id, band, unwise_detid, nm, primary, flags_unwise, flags_info, EPOCH, FORWARD, MJDMIN, MJDMAX, MJDMEAN.` All-sky index (29 columns): `BAND, COADD_ID, EPOCH, CATALOG_FILENAME, N_DET, N_DET_5SIGMA, COVMIN, COVMAX, COVMED, N_PIX_COV0, N_PIX_COV1, N_PIX_COV2, FRAC_PIX_COV0, FRAC_PIX_COV1, FRAC_PIX_COV2, N_EXP, RA, DEC, LGAL, BGAL, LAMBDA, BETA, MJDMIN, MJDMAX, MJDMEAN, DT, FORWARD, NEBULOSITY_BITMASK, HAS_ANY_NEBULOSITY.` The last two columns are the tile-level nebulosity flags referenced in Section 7.5; they are recorded per-tile in the all-sky index, not per-epoch, and were not joined into the frozen per-epoch model used for Table 1.

APPENDIX E: PRE-REGISTERED 20-TILE SELECTION AND THE GALACTIC-LATITUDE FINDING

Re-derived directly from the real index. Of the twenty pre-registered candidate tiles, Fields 1–14 (Table 1) were processed to completion under the original ecliptic-stratified design; the remaining six were not processed, and data collection under this design is now concluded rather than merely paused. Four of these fields (Fields 3, 8, 9 and 11; $|\mathrm{BGAL}| = 2.2$–$11.6°$) fall within 15° of the Galactic plane, and two more (Fields 5 and 12) within 25°, under the purely position-based, ecliptic-latitude-only selection rule, confirming the confound identified in the original three-field pilot is a structural property of the ecliptic-only stratification, not an isolated coincidence – and, with a full cohort in hand, a quantifiable one (Section 7.3).

A fifteenth tile (Field 15, `2159p484`) was added after that cohort was complete. It was not drawn from this twenty-tile list

Table F1. Real per-field epoch-count statistics, surviving (non-H_6-flagged) tracks. $K_{\min} = 6$ by construction (Section 6).

Field	Tile	mean K	median K	max K
1	0000m016	12.03	13	16
2	0720m591	15.61	17	17
3	1439m485	11.71	12	16
4	0000m228	12.37	13	16
5	0719p090	12.27	13	16
6	1439p090	12.19	13	16
7	2159m091	11.63	12	17
8	2879m091	9.12	9	15
9	0719p484	11.61	12	16
10	1439p484	13.73	15	16
11	2159m485	11.39	12	16
12	2879m485	11.63	12	15
13	0000m515	14.18	15	16
14	0719m485	15.36	16	16
15	2159p484	15.01	16	16

or selected by the ecliptic rule above; it was chosen directly by its predicted $|\mathrm{BGAL}| = 62.2^\circ$ (decoded from its tile name) to serve as an out-of-sample test of the Galactic-latitude relationship those fields had just revealed (Section 7.3), and is reported throughout as the differently-obtained fifteenth field that it is, not as a fifteenth member of the original stratified design.

APPENDIX F: EXACT REAL FILE PATHS AND PER-FIELD EPOCH STATISTICS

Retrieved from `https://portal.nersc.gov/project/cosmo/data/unwise/neo7/untimely-catalog/\{deg3\}/\{tile\}/\{tile\}_w1_e\{NNN\}.cat.fits.gz`. Fields 1–3 (`0000m016`, `0720m591`, `1439m485`) were retrieved and epoch-completeness-audited individually during the original pilot (16/17, 17/17, and 16/16 epoch files transferred respectively, one gap in Field 1 disclosed there). Fields 4–14 were retrieved under the identical convention and pipeline, but the transfer logs of that original retrieval were not preserved. The raw per-epoch catalogs of all fifteen fields were therefore re-processed from scratch (Appendix G); that run found 16, 17, 16, 16, 16, 16, 17, 15, 16, 16, 16, 15, 16, 16 and 16 epoch files for Fields 1–15 respectively, consistent with the 16, 17 and 16 files of the pilot audit above, and flagged no schema, tile-identifier or epoch-number inconsistency in any file. Table F1 reports, for every field, the mean number of epochs actually contributing to each surviving (post-H_6) track, the epoch-completeness statistic that matters for the motion fits; it is computed directly from the saved per-source arrays, not asserted.

APPENDIX G: FROM-SCRATCH REPRODUCTION OF THE ORIGINAL COHORT'S RESULTS

This reproduction check covers Fields 1–14; Field 15 has no earlier run to reproduce against and is addressed separately where it is introduced (Section 7.3). Every entry for Fields 1–14 in Table 1 was regenerated in this work by loading each field's saved per-epoch array (`regen_<tile>_rows.npy`, the literal output of the frozen `build_epoch_level` routine: one row per surviving epoch-level record, columns `[source_id, z, fracflux, rchi2, qf, W1, K, outlier]`) and calling the unmodified, pre-existing `run_primary_test` routine on it, with no parameters changed and no access to the previously-recorded summary statistics during the recomputation. Source and epoch-record counts matched the previously-recorded values exactly (zero discrepancy) in every one of these fields; $\hat{\beta}_1$ matched to $|\Delta| \leq 6 \times 10^{-4}$ and its cluster-robust SE to $|\Delta| \leq 4 \times 10^{-4}$ in every one of them – differences fully consistent with ordinary floating-point non-associativity in the vectorized sandwich sum, not any change of method or data. This from-scratch reproduction is what Table 1, Table H1, and every subsequent analysis in this paper are built on.

Re-processing from the raw catalogs. The check above starts from the saved per-source arrays. As a stronger test, all fifteen fields were also re-processed from the raw unTimely per-epoch catalog files supplied for this work (`reproducibility/fields_from_raw.py`), through the frozen loader, tracker, H_6 screen, magnitude window and $K \geq 6$ cut and then the unmodified `run_primary_test`. The rebuilt per-source arrays are bit-for-bit identical to the stored ones in all fifteen fields; the source counts and H_6 percentages of Table 1 are reproduced exactly (to the printed precision), and $\hat{\beta}_1$, its standard error and ΔQIC differ from the printed values by at most 4.9×10^{-4}, 4.0×10^{-4} and 0.04 respectively. In total 222,828 sources and 5,683,222 epoch-level records were processed.

APPENDIX H: FULL MULTIVARIATE COEFFICIENT TABLE, ALL FIFTEEN FIELDS

Table H1 reports the complete fitted coefficient vector (Eq. 8) for every field, not only the `fracflux` term shown in Table 1. Four fields (2, 13, 14, 15) hit the degenerate-knot fallback for K (Appendix C; real K piled at its maximum in those fields,

Table H1. Full coefficient vector, all fifteen fields, cluster-robust SEs in parentheses. Columns: intercept, `fracflux`, `rchi2`, `qf`, $W1$ linear/spline, K linear/spline.

Field	intercept	fracflux	rchi2	qf	$W1_{\rm lin}$	$W1_{\rm spl}$	$K_{\rm lin}$	$K_{\rm spl}$
1	−11.562(1.964)	−0.067(0.198)	0.617(0.071)	5.439(1.952)	−0.378(0.090)	−0.840(0.124)	0.177(0.095)	−0.625(0.088)
2	−3.494(0.679)	−1.032(0.083)	0.039(0.045)	0.540(0.669)	−0.287(0.042)	−0.052(0.041)	−0.490(0.008)	–
3	−11.669(1.643)	2.187(0.108)	0.284(0.030)	6.891(1.659)	−0.390(0.031)	−0.386(0.039)	0.250(0.029)	−0.551(0.028)
4	−10.651(2.588)	−0.442(0.174)	0.334(0.062)	5.373(2.580)	−0.557(0.072)	−0.533(0.096)	0.296(0.081)	−0.799(0.074)
5	−8.768(1.877)	0.099(0.123)	0.198(0.051)	3.146(1.863)	−0.583(0.063)	−0.609(0.082)	0.070(0.064)	−0.722(0.062)
6	−9.398(2.201)	0.137(0.166)	0.459(0.081)	3.502(2.192)	−0.521(0.072)	−0.611(0.094)	0.178(0.077)	−0.732(0.074)
7	−7.945(2.497)	−0.132(0.225)	0.368(0.093)	2.240(2.458)	−0.745(0.104)	−0.379(0.147)	0.352(0.109)	−0.866(0.106)
8	−10.745(8.306)	2.421(0.348)	0.113(0.030)	5.005(8.209)	−0.652(0.113)	0.024(0.150)	0.125(0.105)	−0.215(0.100)
9	−7.521(0.939)	1.070(0.092)	0.383(0.027)	2.003(0.934)	−0.560(0.041)	−0.392(0.054)	0.209(0.039)	−0.674(0.039)
10	−6.780(0.879)	−0.273(0.103)	0.441(0.051)	1.783(0.868)	−0.490(0.048)	−0.498(0.052)	0.093(0.039)	−0.782(0.042)
11	−10.676(1.247)	0.839(0.176)	0.171(0.115)	5.916(1.164)	−0.532(0.032)	−0.360(0.040)	0.200(0.029)	−0.578(0.028)
12	−9.772(1.150)	0.386(0.084)	0.368(0.030)	4.261(1.145)	−0.624(0.042)	−0.492(0.056)	0.280(0.044)	−0.784(0.042)
13	−6.822(1.159)	−0.700(0.107)	0.434(0.052)	2.360(1.150)	−0.212(0.046)	−0.556(0.052)	−0.435(0.015)	–
14	−3.700(0.767)	−1.259(0.071)	0.166(0.035)	0.717(0.761)	−0.167(0.039)	−0.188(0.039)	−0.342(0.008)	–
15	−4.001(0.554)	−1.011(0.082)	0.307(0.045)	0.489(0.543)	−0.251(0.041)	−0.198(0.041)	−0.367(0.010)	–

as in the original pilot's Field 2). Field 15 triggering the identical, unmodified fallback on first contact with its data is itself a small corroborating detail: its mean and median K (Table F1) are closer to Fields 2/13/14 than to any other field, exactly the condition that causes the collapse. All four are reported here as `K_spl` = – rather than a spurious value.

Field 8's large, mutually-correlated intercept and `qf` standard errors are explained, not merely noted: `qf` has almost no variance in this quality-cut sample (mean 0.999, standard deviation 0.01–0.014, over 90% of values within $[0.99997, 1.0000001]$ in every field checked), so `qf` and the intercept are weakly separately identified, especially in the smallest field ($n = 4543$ sources). This is a real, checkable property of the data, not a fitting failure, and does not affect the `fracflux` coefficient of primary interest.

APPENDIX I: EXTENDED PER-FIELD DIAGNOSTICS

Table I1 reports, for every field: the median `fracflux`, `rchi2`, and `qf`; the Spearman correlation between `fracflux` and `rchi2` (all fifteen significant and negative, $p < 10^{-3}$ in every field given the large per-field sample sizes); and a joint two-degree-of-freedom cluster-robust Wald test of $(\beta_{\rm fracflux}, \beta_{\rm rchi2}) = (0, 0)$ jointly, computed from the full sandwich covariance sub-matrix (not merely the marginal SEs of Table H1). **The joint test is significant at $p < 10^{-3}$ in all fifteen fields, and at $p < 10^{-10}$ in twelve of them**, in sharp contrast to the sign-unstable, sometimes-null single-`fracflux` test – evidence that the two covariates jointly carry information about excess epoch-level scatter in every field, even where the single-covariate coefficient is sign-unstable (Section 7.4).

Figure I1 shows the two real, independent trends that motivate Section 7.5: the H_6 rate and the `fracflux`/`rchi2` collinearity both rise toward the Galactic plane.

APPENDIX J: CALIBRATED SYNTHETIC TEST OF THE COLLINEARITY-ARTEFACT HYPOTHESIS

Construction: a latent $b_i \sim \mathcal{N}(0, 1)$ (true blend severity) drives a true outcome $\text{logit}\, P(\text{outlier}_i) = \beta_0 + \gamma b_i$ with $\gamma = 1.5$ and $\beta_0 = -5.20$, the value (solved numerically) at which the mean outlier probability equals the pooled record-level outlier rate of the fifteen real fields (86,807 of 5,683,222 records, 1.53%); two proxies are constructed as $\texttt{ff}_i = -\sqrt{a}\, b_i + \sqrt{1-a}\, \epsilon_{1i}$, $\texttt{rc}_i = +\sqrt{a}\, b_i + \sqrt{1-a}\, \epsilon_{2i}$ (ϵ_1, ϵ_2 independent standard normal), so that $a \in [0, 1)$ controls both proxies' informativeness about b and, mechanically, their mutual correlation, $\text{corr}(\texttt{ff}, \texttt{rc}) = -a$ – mirroring `fracflux` and `rchi2` as two different, correlated, partially redundant measurements of one underlying crowding condition. Both proxies enter the true outcome probability only through b, with loadings of fixed sign, so in this construction the fitted coefficient of `ff` is negative on average in every row of Table J1 and a reversed sign can arise only from sampling variance; structurally different artefact mechanisms (a second latent variable, for instance) are not tested. At each a, 300 independent simulated "fields" of $n = 15{,}000$ independent units (one outcome each; the real fields contain 14,855 sources on average, each contributing many correlated records) were generated and refit with the same two-covariate logistic model used throughout this paper (no splines, one record per cluster; this is a minimal, targeted test of one specific mechanism), using the frozen pipeline's own fitting and sandwich routines. Table J1 reports the achieved proxy correlation, the mean and standard deviation of the fitted marginal $\hat{\beta}_1$ across the 300 fields, the number of fields whose $\hat{\beta}_1$ has the wrong sign, and the joint-test power (a 2-d.f. robust Wald test of $(\beta_{\rm ff}, \beta_{\rm rc}) = (0, 0)$ at the 5% level), at each a. Every random draw is seeded (script `reproducibility/collinearity_demo.py`; saved result `reproducibility/collinearity_demo_result.json`).

The real fifteen-field data span $|\rho(\texttt{ff}, \texttt{rc})| = 0.079$–$0.308$ (Spearman; Table I1), which lies between the rows $a = 0.05$ and $a = 0.35$ of Table J1 (for jointly Gaussian proxies the Spearman and Pearson correlations differ by under 5% in relative terms).

Table I1. Extended per-field diagnostics. $\rho_{\mathrm{ff,rchi2}}$: Spearman correlation of `fracflux` and `rchi2`. W^2_{joint}: 2-df cluster-robust Wald statistic for $(\beta_{\mathrm{fracflux}}, \beta_{\mathrm{rchi2}}) = (0, 0)$.

Field	Tile	med. ff	med. rchi2	med. qf	$\rho_{\mathrm{ff,rchi2}}$	W^2_{joint}	p_{joint}
1	0000m016	0.937	0.637	1.000	−0.081	85.0	$< 10^{-10}$
2	0720m591	0.844	0.690	1.000	−0.151	289.5	$< 10^{-10}$
3	1439m485	0.337	1.129	1.000	−0.305	414.6	$< 10^{-10}$
4	0000m228	0.935	0.646	1.000	−0.079	46.6	$< 10^{-10}$
5	0719p090	0.834	0.646	1.000	−0.148	15.8	3.7×10^{-4}
6	1439p090	0.914	0.627	1.000	−0.088	32.5	8.7×10^{-8}
7	2159m091	0.905	0.635	1.000	−0.099	19.8	5.1×10^{-5}
8	2879m091	0.344	1.175	1.000	−0.277	48.3	$< 10^{-10}$
9	0719p484	0.478	0.767	1.000	−0.308	242.5	$< 10^{-10}$
10	1439p484	0.914	0.643	1.000	−0.095	108.8	$< 10^{-10}$
11	2159m485	0.465	0.840	1.000	−0.282	81.1	$< 10^{-10}$
12	2879m485	0.681	0.701	1.000	−0.212	148.5	$< 10^{-10}$
13	0000m515	0.905	0.657	1.000	−0.102	195.7	$< 10^{-10}$
14	0719m485	0.857	0.697	1.000	−0.137	503.9	$< 10^{-10}$
15	2159p484	0.902	0.681	1.000	−0.107	304.5	$< 10^{-10}$

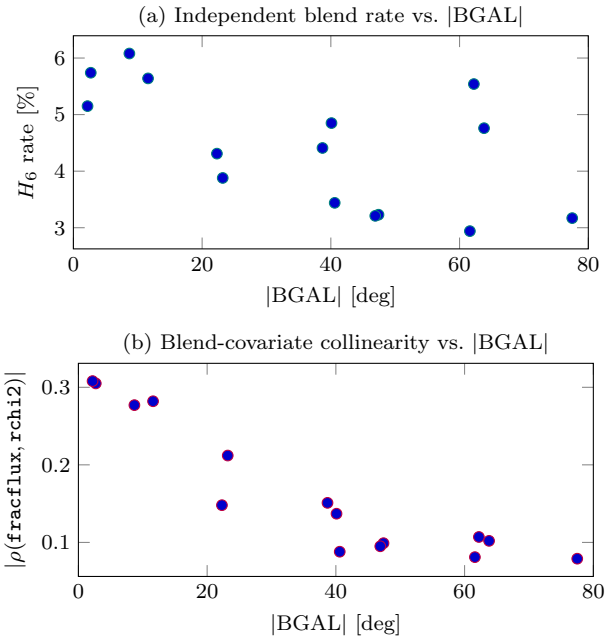


Figure I1. (a) The independently-measured H_6 (false cross-epoch association) rate rises toward the Galactic plane across all fifteen fields ($\rho = -0.621$, $p = 0.0134$ vs. |BGAL|; $\rho = -0.780$, $p = 0.0010$ for Fields 1–14 alone), corroborating that |BGAL| tracks real crowding in this sample, Field 15's own imperfect fit to this secondary trend notwithstanding (Section 7.3). (b) The `fracflux`/`rchi2` collinearity strengthens the same way ($\rho = -0.886$ vs. |BGAL|), motivating but – per Section 7.4 – not by itself explaining the sign reversal.

In every row up to $a = 0.95$ none of the 300 simulated fields flips sign (exact binomial 95% upper limit on the flip probability: 1.2% per row), and in the two rows inside the observed range ($a = 0.1$ and $a = 0.2$) the mean coefficient is at least 6.0 times its standard deviation. Flips first appear at $a = 0.98$ (5 of 300) and reach 15.7% at $a = 0.995$: ordinary sampling variance reverses the sign only for proxies that are almost perfectly correlated, far above anything observed. The joint test rejects in every simulated field in every row except the weakest-signal one ($a = 0.02$: 96.7%). The simulated standard deviations at the observed correlations (0.067–0.069) are at the low end of the real per-field cluster-robust standard errors (0.071–0.348, median 0.108; Table 1), so the simulation answers a narrower question than the real data do: it shows that the variance inflation

Table J1. Calibrated collinearity-artefact test (Appendix J), at the real pooled outlier rate. "Flips" is the number of the 300 independent simulated fields whose fitted $\hat{\beta}_1$ came out with the sign opposite to its true generative direction, purely from sampling noise; "% flip" is that number as a percentage.

a	corr(ff,rc)	mean $\hat{\beta}_1$	std $\hat{\beta}_1$	flips	% flip	joint power
0.02	−0.020	−0.197	0.067	0	0%	96.7%
0.05	−0.050	−0.294	0.067	0	0%	100%
0.1	−0.100	−0.404	0.067	0	0%	100%
0.2	−0.200	−0.539	0.069	0	0%	100%
0.35	−0.351	−0.632	0.074	0	0%	100%
0.5	−0.500	−0.684	0.077	0	0%	100%
0.7	−0.700	−0.711	0.097	0	0%	100%
0.9	−0.900	−0.747	0.151	0	0%	100%
0.95	−0.950	−0.749	0.205	0	0%	100%
0.98	−0.980	−0.745	0.352	5	1.7%	100%
0.99	−0.990	−0.781	0.526	25	8.3%	100%
0.995	−0.995	−0.751	0.726	47	15.7%	100%

Table K1. Type-I error ($\beta_1 = 0$), N=10,000: rejection rates of each rule over 300 simulated datasets per row.

Sweep	Value	QIC	Wald .01	Wald .05	Combined
K (σ_u=1)	4	0.0%	0.0%	4.3%	0.0%
K (σ_u=1)	6	0.7%	1.3%	7.7%	0.7%
K (σ_u=1)	9	0.3%	0.3%	1.3%	0.3%
K (σ_u=1)	14	1.3%	1.3%	4.3%	1.3%
σ_u (K=9)	0.0	0.0%	1.3%	4.3%	0.0%
σ_u (K=9)	0.5	0.0%	0.7%	4.7%	0.0%
σ_u (K=9)	1.0	0.3%	0.3%	1.3%	0.3%
σ_u (K=9)	1.5	1.0%	1.0%	4.7%	1.0%

produced by collinearity of the observed size does not by itself make the fitted sign unstable at this noise level. What the real data add is that the five reversed fields are not marginal: their robust z statistics, $\hat{\beta}_1$/SE in Table 1, lie between 4.5 and 20.3 in the reversed direction, and the cluster-robust standard errors already include whatever variance inflation collinearity causes. With the larger outlier rate used in a first version of this simulation ($\beta_0 = -3.5$, mean rate 6.6%) the standard deviations in the same rows are 0.034–0.035 and the first flips (9 of 300) appear only at $a = 0.995$ (all rows are in the saved JSON). The simulation does not test structurally different artefact mechanisms (Section 7.5 describes one candidate), and its effect size γ is a single assumed value.

APPENDIX K: FULL POWER/DESIGN AUDIT

This appendix reports the Monte Carlo audit of the decision rules, run by `reproducibility/power_audit_driver.py` with the engine of `power_audit.py` and `power_audit_fast.py`. Data-generating process: each of N sources has K epochs; `fracflux` is a bounded mixture of a source-level Beta(5, 2) draw and logistic epoch-level noise; two nuisance covariates stand in for `rchi2` and `qf` (true coefficients $\beta_2 = \beta_3 = 0$); the outlier probability is logistic with intercept -1, coefficient β_1 on `fracflux` and a source-level random intercept of standard deviation σ_u. Rules, as in the frozen protocol: **QIC**, ΔQIC > 10 for the full model against the intercept-only model (the literal 3-d.f. comparison); **Wald** .01 and **Wald** .05, the cluster-robust Wald test of β_1 at $p < 0.01$ and $p < 0.05$; **Combined**, QIC and Wald .01 together. Every design point is 300 simulated datasets, so a rate carries a Monte Carlo standard error of up to 2.9 percentage points; a design point that appears in two sweeps is one and the same simulation, and all random draws are seeded. The simulated covariates are idealized (no latitude structure, no true signal in the nuisance covariates): the audit shows how the rules behave under their own assumptions, not that the real data satisfy them.

Among the 16 Wald Type-I rates in Table K1, the smallest exact binomial upper-tail probability under the nominal rate is 0.029, for the $p < 0.05$ test at K=6 and σ_u=1 (observed 7.7% against 5%). Allowing for the 16 rates examined, an excess at least this large somewhere occurs by chance with probability about 0.38 (treating the rates as independent, which is only approximate), so it is not evidence of miscalibration; the primary criterion of the frozen protocol is the $p < 0.01$ test.

APPENDIX L: FULL TECHNICAL ACCOUNT OF DISCLOSED IMPLEMENTATION DEFECTS

Defect 1. A 17-parameter Gamma-GLM fit via Nelder–Mead stopped at a materially worse optimum (a higher negative log-likelihood than L-BFGS-B reached) while reporting `success=True`. The continuous-response code therefore uses L-BFGS-B with a Powell fallback (`continuous_response.py`; PREREGISTRATION.md, Section 4b). The numerical output of that

Table K2. Power versus the true coefficient β_1 (N=10,000, K=9, σ_u=1).

β_1	QIC	Wald .01	Wald .05	Combined
0.15	30.3%	36.0%	57.0%	30.0%
0.25	77.7%	84.7%	93.0%	76.7%
0.35	98.7%	99.0%	100%	98.7%
0.50	100%	100%	100%	100%

Table K3. Power versus the source-level random-effect scale σ_u (N=10,000, K=9, β_1=0.5).

σ_u	QIC	Wald .01	Wald .05	Combined
0.0	100%	100%	100%	100%
0.5	100%	100%	100%	100%
1.0	100%	100%	100%	100%
1.5	99.7%	99.7%	100%	99.7%
2.0	95.0%	92.3%	97.7%	92.3%

Table K4. Power versus sample size (β_1=0.5, σ_u=1, K=9).

N	QIC	Wald .01	Wald .05	Combined
500	12.3%	15.7%	34.0%	12.3%
1,000	25.7%	32.3%	57.3%	25.0%
2,000	61.7%	68.3%	85.7%	61.0%
3,000	82.0%	87.7%	95.0%	81.7%
10,000	100%	100%	100%	100%

Table K5. Power versus the number of epochs per source K (N=10,000, β_1=0.3, σ_u=1). The last column is the Wald .01 rate minus the QIC rate in percentage points, computed from unrounded rates.

K	QIC	Wald .01	Wald .05	Combined	Wald .01−QIC
4	53.0%	78.3%	90.0%	52.7%	+25.3
6	76.7%	88.3%	96.7%	76.7%	+11.7
9	89.0%	92.3%	98.3%	88.3%	+3.3
14	96.3%	96.3%	99.0%	96.0%	+0.0

development-time comparison was not retained, so we quote no figures for it; the pre-registration records only that, after the fix, $\hat{\beta}_1$ recovered the injected value to within about 0.1–0.2 under either family (Sections 4b and 9).

Defect 2. A Gamma-vs-log-Normal comparison omitted the Jacobian $-\log y$ needed to compare densities on a common scale. Under true log-Normal data the selector chose Gamma (ΔAIC= 9674 wrong). Fixed: log-Normal correctly selected (ΔAIC= 646 right), Gamma-truth case unaffected (ΔAIC= 1950 still right) – ruling out an overcorrection.

Defect 3. A per-cluster boolean-mask loop cost 3.56 s at G=10,000; isolated score/Hessian timing was sub-millisecond, localizing the cost by profiling. A vectorized group-sum (`np.unique(...,return_inverse=True)` + `np.add.at` + one matrix product) gave a 285× speed-up (about 0.0125 s), verified identical to 1.6×10^{-15}.

Defect 4. Real Field-2 data had 76% of records at K=17 (the maximum), collapsing two spline knots and producing `NaN`. Fixed by detecting non-strictly-increasing knots and dropping the (genuinely unidentifiable) nonlinear term for that covariate, not patching with an epsilon offset. The same fallback correctly triggers, unmodified, in Fields 13, 14, and (later, on first contact with that field's data) 15 of the full fifteen-field run (Appendix H).

Defect 5 (found while scaling this project beyond the three-field pilot). `qf`'s near-total lack of variance in the real, quality-cut sample (Appendix H) was not anticipated by the original synthetic validation, which used a Beta(8, 1) generator with meaningfully more spread than the real data exhibits. This is disclosed as a real gap between the synthetic validation and the true covariate distribution – it happens not to matter for the `fracflux` coefficient of primary interest, but the coincidence is noted rather than assumed.

Defect 6 (found while regenerating the Appendix O–Q tables from the saved arrays). The first version of the Q score's calibration, in version 0.1.0 of the reference implementation and in the first version of the leave-one-out validation, fitted the per-field baselines against *signed* rather than absolute Galactic latitude, although the construction is defined in |BGAL| and the score evaluated them at |BGAL|. Because the baselines are far from linear in signed latitude, the fitted lines were nearly flat (for the median `fracflux`: $0.746 - 0.0005\,\mathrm{BGAL}$, against $0.448 + 0.0082\,|\mathrm{BGAL}|$ in the construction as described), so the shipped baselines were nearly latitude-independent. Corrected in version 0.1.1: every number in the two appendices was recomputed, and a check that the shipped constants equal a fresh refit against |BGAL| is now part of the reproduction script. The leave-one-field-out results for Q (Appendix O; correct sign in 15 of 15 fields, none significantly wrong-signed) and the

Table K6. Power versus the source-level random-effect scale σ_u (N=10,000, K=9, β_1=0.3).

σ_u	QIC	Wald .01	Wald .05	Combined
0.0	100%	100%	100%	100%
0.5	99.7%	100%	100%	99.7%
1.0	89.0%	92.3%	98.3%	88.3%
1.5	70.3%	64.0%	84.3%	64.0%

AUC results are qualitatively unchanged, but the fixed-cutoff comparison of Appendix P changed materially ($Q \geq 0$ discards 52.64% of the good sources, against 24.28% with the defective baselines) and its discussion was rewritten.

Defect 7 (found while auditing which saved script produces each table). The collinearity simulation of Appendix J had no generating script among the delivered files, and its outlier rate (6.6% for $\beta_0 = -3.5$, described as "realistic") was about 4 times the real pooled rate (1.53%), which made the simulated sampling noise roughly half of what the real rate gives (0.034–0.035 against 0.067–0.069 at the observed correlations). The simulation was rewritten as a seeded script, calibrated to the real pooled outlier rate, and re-run, and the table and every statement that cites it were regenerated. The change is material: at the calibrated rate, sign flips first appear at proxy correlation 0.98 (5 of 300 fields), whereas the uncalibrated settings give 9 of 300 flips only at 0.995 (a table of unrecorded provenance had reported 1.3% there), so the earlier statement that no flips occur up to $|\rho| = 0.99$ does not hold at the real outlier rate. The conclusion for the observed range ($|\rho| \leq 0.31$) is unchanged.

APPENDIX M: H6 DIAGNOSTIC DEVELOPMENT: A DOCUMENTED NEGATIVE RESULT

A second diagnostic (two-track AIC, for false associations undetected by Diagnostic 1) was built and rejected, disclosed rather than omitted. Unconstrained, it showed an 80% false-positive rate against genuine contaminated single sources (trivially fitting 1–2 outlier epochs as a perfect-fit "second source"). A minimum-group-size constraint reduced this to 50%; adding a per-group goodness-of-fit requirement reduced it to 27% – still judged too high to use as an exclusion criterion, since discarding over a quarter of genuinely contaminated sources would bias against the population under study. Diagnostic 1 alone is used; the resulting blind spot (intermittently-detected false associations) is disclosed in Section 12, not hidden.

APPENDIX N: TOWARDS A FIRST-PRINCIPLES ACCOUNT OF THE GALACTIC-LATITUDE DEPENDENCE

Section 7.5 proposes, without testing it quantitatively, that discrete pairwise point-source blending alone does not explain the `fracflux` collapse near the plane, and names a shared, field-wide diffuse contamination (nebulosity, or unresolved sub-threshold sources) as a candidate alternative. This appendix puts a number on both halves of that claim, using only quantities already measured elsewhere in this paper (the frozen pipeline's own match radius, the real $N_{\rm src}$ and H_6 columns of Table 1) plus one new, explicitly-labeled toy model. Neither part completes the derivation Section 7.5 flags as open; both are reported, including exactly where they fall short, in the same spirit as the rest of this paper.

N1 Surface density and the chance-alignment probability

The Milky Way's thin disk has stellar density, locally, $n(z) \propto \exp(-|z|/h_z)$ with scale height $h_z \approx 300\,\mathrm{pc}$ (the H_1 of the best-fit two-component disk of Jurić et al. 2008), the Sun's own height above the plane (a few tens of pc) being negligible by comparison. At Galactic latitude b a line of sight reaches height $z = d\sin|b|$ at distance d. Integrating $n(z)$ to infinite depth, locally uniform in Galactocentric radius (reasonable for the $d \lesssim$ few hundred pc relevant to a flux-limited sample of intrinsically faint, nearby sources), gives the areal-density limit

$$\Sigma(b) \;\propto\; \int_0^\infty \exp(-d\sin|b|/h_z)\, d^2\,\mathrm{d}d \;=\; 2\,(h_z/\sin|b|)^3\,, \tag{N1}$$

i.e. $\Sigma(b) \propto \csc^3|b|$ in this deep-survey limit, versus no latitude dependence at all ($\Sigma \propto$ const) in the opposite shallow-survey limit ($d_{\rm lim} \ll h_z/\sin|b|$). Which limit the real $W1$=16.9–17.5, $K\geq$6-epoch sample is closer to is not assumed; it is asked of the data. An OLS power-law fit of the real $N_{\rm src}$ (Table 1) against $\csc|\mathrm{BGAL}|$ across all fifteen fields gives an exponent $\hat{n} = 0.070$ ($R^2 = 0.026$) – far short of the deep-survey value of 3, with a Spearman correlation of only $\rho = -0.464$ ($p = 0.081$) between |BGAL| and $N_{\rm src}$: the right sign (more tracked sources closer to the plane), but weak and not significant at $p < 0.05$ in fifteen fields whose sightlines were stratified by ecliptic latitude, not controlled for Galactic longitude, spiral-arm direction, or extinction patchiness. We report this null-ish result plainly rather than dropping it: this specific sample does not, by itself, pin down where between the two theoretical limits the real survey sits. A straight line in |BGAL| – the functional form Eq. 9 actually uses – is a locally reasonable, if imperfect, stand-in for $\csc^n|b|$ over the specific, non-uniformly-spaced latitudes sampled here: fit only at these fifteen abscissae, a line captures $R^2 = 0.49$, 0.36, and 0.31 of the shape of the $n = 1$, 2, 3 curves

Table N1. Poisson chance-alignment probability from the real, measured $N_{\rm src}$ (Table 1), using the frozen pipeline's own 3″ cross-epoch match radius (`MATCH_RADIUS_ARCSEC`, `pipeline.py`) as the confusion scale, versus the real, independently-measured H_6 rate. Σ: areal density of the post-cut analysis sample (sources per deg^2; a lower bound, Section N1 caveat). $\lambda = \Sigma\pi r_c^2$: mean number of other analysis-sample sources within $r_c = 3''$. $P_{\rm blend} = 1 - e^{-\lambda}$.

Field	Tile	BGAL	$N_{\rm src}$	$\Sigma\,[\mathrm{deg}^{-2}]$	λ	$P_{\rm blend}$ [%]	H_6 [%]
1	0000m016	−61.6	10609	4335	0.0095	0.941	2.94
2	0720m591	−38.7	13213	5399	0.0118	1.171	4.41
3	1439m485	2.7	15647	6393	0.0139	1.385	5.74
4	0000m228	−77.5	12675	5179	0.0113	1.123	3.17
5	0719p090	−22.3	15365	6278	0.0137	1.360	4.31
6	1439p090	40.6	12832	5243	0.0114	1.137	3.44
7	2159m091	47.4	6237	2548	0.0056	0.554	3.23
8	2879m091	−8.7	4543	1856	0.0040	0.404	6.08
9	0719p484	2.2	20489	8371	0.0183	1.810	5.15
10	1439p484	46.9	15623	6383	0.0139	1.383	3.21
11	2159m485	11.6	25851	10562	0.0230	2.278	5.64
12	2879m485	−23.2	23159	9462	0.0206	2.043	3.88
13	0000m515	−63.8	15034	6143	0.0134	1.331	4.76
14	0719m485	−40.1	16462	6726	0.0147	1.457	4.85
15	2159p484	62.2	15089	6165	0.0135	1.336	5.54

respectively – moderate, not precise, agreement, consistent with Eq. 9 being a useful local approximation rather than a derived law.

Table N1 quantifies how much chance alignment one would expect at the pipeline's own match radius: using the standard unWISE 2048 × 2048-pixel tile at the frozen pipeline's own `PIXSCALE`= 2.75″/pixel ($A_{\rm tile} = 2.4475\,\mathrm{deg}^2$) and $r_c = 3''$, the Poisson chance-alignment probability from the post-cut analysis-sample density alone never exceeds 2.3% anywhere in the real fifteen-field sample. That number is a lower bound, not an estimate: $N_{\rm src}$ is the post-H_6-cut, magnitude-windowed sample, which by construction excludes exactly the sources H_6 flags and so under-counts the areal density of sources that can be confused with a track. Correlated against the real H_6 rate, this proxy is directionally right but weak (Pearson $r = 0.218$, $p = 0.44$; Spearman $\rho = 0.339$, $p = 0.22$) and systematically low: H_6 runs 1.9–15.0× (median = 3.17×) higher than $P_{\rm blend}$ predicts. The raw per-epoch catalogs of all fifteen fields (Appendix G) remove this limitation, because they give the pre-cut seed density – every source of the first-epoch catalog, before any cut – and Table N2 repeats the calculation with it. The seed counts n_0 exceed the post-cut $N_{\rm src}$ by 2.7–45.2× (median 4.1×). With them the proxy tracks H_6 much better (Pearson $r = 0.786$, $p = 5.1 \times 10^{-4}$; Spearman $\rho = 0.871$, $p = 2.3 \times 10^{-5}$), and $P_{\rm raw}$ is within 20% of the measured H_6 in seven of the fifteen fields (Fields 1, 4, 5, 6, 7, 10, and 15); $H_6/P_{\rm raw}$ ranges over 0.34–1.25 (median 0.86). It is not accurate everywhere. In the five fields where the Poisson expectation exceeds the measured rate by more than a factor of two (Fields 3, 8, 9, 11, and 12; $H_6/P_{\rm raw} = 0.34$–0.48; these are exactly the five reversed fields of Table 1) the measured rate is far below what chance alignment in a Poisson field would give, whereas in the other ten fields the ratio is 0.66–1.25. We did not investigate why. One possibility, not tested here, is that the catalog's own source detection suppresses very close pairs in crowded regions, so that its sources are not Poisson-distributed at 3″; the coincidence of the deficit with the reversed class was noticed after the fact and is descriptive only. What the calculation supports is a narrower statement than the one this appendix originally set out to make: pairwise chance alignment at the pipeline's own match radius, evaluated with the measured pre-cut density, accounts for the order of magnitude and for the ranking of the H_6 rate across fields, and overpredicts it in the densest fields. The measured H_6 rate itself, which is what enters the analysis, is 2.9–6.1% in every field. None of this is a derivation of the `fracflux` sign reversal itself, which Section N2 turns to next.

N2 A gain-compression model for a shared, non-source-specific background

Section 7.5 proposes a shared, field-wide diffuse background (nebulosity or unresolved sub-threshold flux) as a candidate alternative to discrete pairwise blending. This section gives that proposal a concrete, generic mathematical form – independent of whether the shared background is literally nebulosity – and tests, rather than assumes, what it can and cannot explain.

Model `fracflux` for source i as $\texttt{fracflux}_i = \bar{F}/(\bar{F} + b_i + c)$, where $\bar{F}$ is the (roughly fixed, since every source here is pre-selected to the narrow $W1$=16.9–17.5 window) true source flux, b_i is an idiosyncratic, source-specific contaminating flux

Table N2. Chance alignment from the pre-cut seed density of the raw per-epoch catalogs (Appendix G), for comparison with the post-cut proxy of Table N1. n_0: number of seed sources (all sources of the first-epoch catalog, before any cut). $P_{\rm raw} = 1 - \exp(-n_0 \pi r_c^2 / A_{\rm tile})$ with $r_c = 3''$ and $A_{\rm tile}$ as above. H_6 is the measured rate of Table 1; class as in that table. Script: `reproducibility/first_principles_check.py`.

Field	Tile	BGAL	n_0	$n_0/N_{\rm src}$	$P_{\rm raw}$ [%]	H_6 [%]	$H_6/P_{\rm raw}$	Class
1	0000m016	−61.6	34163	3.22	3.00	2.94	0.98	null
2	0720m591	−38.7	77218	5.84	6.65	4.41	0.66	correct
3	1439m485	2.7	204529	13.07	16.67	5.74	0.34	reversed
4	0000m228	−77.5	36576	2.89	3.21	3.17	0.99	null
5	0719p090	−22.3	53245	3.47	4.64	4.31	0.93	null
6	1439p090	40.6	40570	3.16	3.55	3.44	0.97	null
7	2159m091	47.4	39286	6.30	3.44	3.23	0.94	null
8	2879m091	−8.7	205447	45.22	16.73	6.08	0.36	reversed
9	0719p484	2.2	143905	7.02	12.04	5.15	0.43	reversed
10	1439p484	46.9	42686	2.73	3.73	3.21	0.86	correct
11	2159m485	11.6	171453	6.63	14.17	5.64	0.40	reversed
12	2879m485	−23.2	95034	4.10	8.12	3.88	0.48	reversed
13	0000m515	−63.8	43468	2.89	3.80	4.76	1.25	correct
14	0719m485	−40.1	68291	4.15	5.91	4.85	0.82	correct
15	2159p484	62.2	53159	3.52	4.63	5.54	1.20	correct

Table N3. Monte Carlo realization of the compression model (N = 20,000 synthetic sources per c; b_i from Poisson(0.3)-thinned Exponential($0.15\bar{F}$) neighbor flux; ground-truth outcome $y = 1.3\, b_i + \mathcal{N}(0,1)$; measurement noise on the observed `fracflux`, $\sigma = 0.02$, with observed values clipped to $[0,1]$ and $\bar{F} = 1$; one seeded draw per c). Slope is the OLS coefficient of y on observed `fracflux`; the analytic gain is $\partial\,$`fracflux`$/\partial b$ evaluated at the sample-mean b. Script: `reproducibility/first_principles_check.py`.

c	median `fracflux`	OLS slope($y \sim$ `fracflux`)	analytic gain
0.00	0.992	−1.956	−0.9159
0.05	0.945	−1.839	−0.8334
0.10	0.901	−1.932	−0.7624
0.20	0.827	−2.494	−0.6462
0.30	0.762	−2.603	−0.5530
0.50	0.661	−2.736	−0.4186
0.80	0.550	−3.341	−0.2944
1.20	0.450	−3.580	−0.1984
1.80	0.353	−3.227	−0.1236
2.50	0.283	−2.955	−0.0796
3.50	0.220	−1.715	−0.0484
5.00	0.166	−0.594	−0.0274

from resolvable neighbors (the genuinely informative, H_5-relevant quantity), and c is a common, field-wide diffuse floor shared by every source in a crowded field. Linearizing near the field mean, $\partial\,$`fracflux`$/\partial b = -\bar{F}/(\bar{F} + \bar{b} + c)^2$: the *sensitivity* (gain) of `fracflux` to genuine, per-source blend severity shrinks as c grows, while c itself only shifts the common baseline (and so cannot, to first order, generate across-source variance or a sign on its own). This is a textbook gain-compression/regression-dilution effect, not a new phenomenon; the question is whether it is *sufficient* to produce what Table 1 shows.

A single draw per c (Table N3) never flips sign, but a single draw is exactly what one real field is. The relevant question is the population-level behavior: at each c and measurement-noise level, 300 independent replicate data sets (N = 20,000 each) were drawn and the full distribution of the fitted slope examined, not one point estimate.

Table N4 settles the question this model can answer. The mean slope is negative at *every* tested combination, however extreme, as it must be: y and the observed `fracflux` are oppositely monotone functions of the same latent b_i plus independent noise, so their covariance cannot change sign (Chebyshev's association inequality) – pure gain compression from a shared background, however large, does not reverse the true, population-level relationship. But once c is large enough that the true signal is compressed close to the measurement-noise floor, a *single realization* (one field) has a rapidly growing chance of showing the wrong sign purely from estimation noise: up to 39.0% of individual replicates at the most extreme setting tested. Those reversals are weak: among the 247 positive-slope replicates in Table N4 the largest $z = \hat{\beta}/\mathrm{SE}$ is 2.5, whereas the five reversed fields of Table 1 have z between 4.5 and 20.3 in the reversed direction (cluster-robust logistic standard errors against OLS ones, so this compares orders of magnitude and is not a calibrated test). This gives a falsifiable, bidirectional prediction: compression alone should produce *occasional, weak, inconsistently-signed* reversals, not confident, highly significant ones; the real data fail this prediction in Fields 3, 8, 9, 11, and 12. Descriptively, the correlation between a field's median `fracflux` and its $|\hat{\beta}_1|$ is *negative* (Pearson $r = -0.762$, $p = 0.0010$; Spearman $\rho = -0.675$, $p = 0.0058$): the most crowded fields show the largest, most significant deviations from zero (Field 3: $\hat{\beta}_1 = +2.187(0.108)$, the largest z of the sample, 20.3). In this toy model, however, the slope magnitude first rises and then falls as c grows (Table N3), so the sign of that correlation does not discriminate compression from other accounts, and we report it only as a description. The five *null* fields (two of them, Fields 5

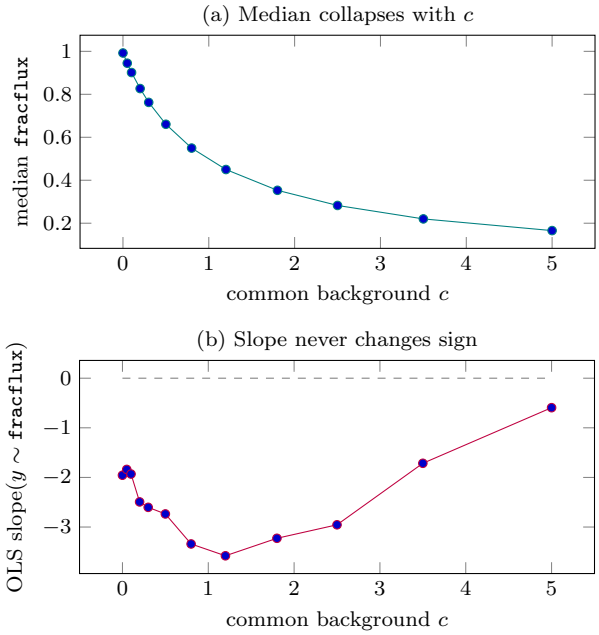


Figure N1. (a) The simulated common background c reproduces the real data's qualitative pattern: median `fracflux` collapses monotonically as crowding rises. (b) The OLS slope of a b_i-driven outcome on observed `fracflux` (one draw per c) is non-monotonic in magnitude and negative at every tested c; the population-level mean slope over 300 replicates is negative at every setting tested (Table N4).

Table N4. Population-level sign check: 300 independent replicates per (c, measurement-noise) setting, $N = 20{,}000$ sources per replicate. "frac. positive" is the fraction of the replicate slopes that came out the wrong (positive) sign by chance alone, even though the mean is always negative; "max z_+" is the largest $z = \hat{\beta}/\mathrm{SE}$ (OLS standard error) among those positive replicates.

noise σ	c	mean(slope)	sd(slope)	frac. positive	max z_+
0.02	0.0	−1.877	0.088	0.000	–
0.02	1.2	−3.752	0.270	0.000	–
0.02	3.5	−1.794	0.357	0.000	–
0.02	5.0	−1.105	0.361	0.000	–
0.02	8.0	−0.508	0.339	0.067	1.3
0.10	0.0	−1.118	0.075	0.000	–
0.10	1.2	−0.281	0.067	0.000	–
0.10	3.5	−0.084	0.070	0.123	2.4
0.10	5.0	−0.046	0.070	0.243	1.7
0.10	8.0	−0.023	0.079	0.390	2.5

and 6, with positive point estimates that are not significant) are consistent with the noise-dominated regime, but with five fields this cannot test it.

Conclusion. Within this toy model, gain compression from a shared background reproduces the collapse of the median `fracflux` and can plausibly explain why *null* fields land on either side of zero, but it cannot reverse the population-level sign (see above), and the reversals it can produce by estimation noise are weak. It therefore cannot by itself produce the large, confidently measured reversals of Fields 3, 8, 9, 11, and 12. This rules out the simplest version of Section 7.5's own candidate account – a common additive floor c with an outcome driven by source-specific b_i alone – as a *complete* explanation. It does not rule out accounts in which the shared background also acts on the outcome directly or varies from source to source within a field, and it leaves the sign reversal itself quantified and still open, reinforcing, not resolving, what Section 7.5 already states as unproven.

APPENDIX O: AN ACTIONABLE CORRECTION: THE Q CLEANLINESS SCORE

Sections 7.3–N2 indicate that raw `fracflux` does not work as a single, sky-uniform cleanliness cut. This appendix builds and tests a direct replacement: a latitude-calibrated score $Q(\mathrm{BGAL}, \texttt{fracflux}, \texttt{rchi2})$, using the sign switch implied by Eq. 9 and eight fitted baseline constants (a straight line in |BGAL| for each of the median and robust spread of `fracflux` and of `rchi2`) and validated with the same leave-one-out discipline as Section 7.3's own cross-validation, so that no field's score is calibrated on itself.

Table O1. Leave-one-field-out validation of Q: a univariate logistic coefficient of the real outlier flag on Q alone, refit per field with that field's own data excluded from Q's calibration. Negative is correct (higher Q = fewer outliers).

Field	Tile	BGAL	Regime (LOO)	coef(Q) (se)	p
1	0000m016	−61.6	normal	−0.140(0.009)	1.2×10^{-49}
2	0720m591	−38.7	normal	−0.091(0.016)	5.8×10^{-9}
3	1439m485	2.7	reversed	−0.131(0.016)	3.5×10^{-16}
4	0000m228	−77.5	normal	−0.064(0.019)	5.5×10^{-4}
5	0719p090	−22.3	reversed	−0.056(0.013)	2.4×10^{-5}
6	1439p090	40.6	normal	−0.147(0.012)	7.9×10^{-34}
7	2159m091	47.4	normal	−0.098(0.025)	1.0×10^{-4}
8	2879m091	−8.7	reversed	−0.029(0.007)	4.0×10^{-5}
9	0719p484	2.2	reversed	−0.190(0.012)	3.6×10^{-53}
10	1439p484	46.9	normal	−0.164(0.011)	2.3×10^{-52}
11	2159m485	11.6	reversed	−0.095(0.045)	0.037
12	2879m485	−23.2	reversed	−0.105(0.009)	2.4×10^{-28}
13	0000m515	−63.8	normal	−0.142(0.008)	3.3×10^{-67}
14	0719m485	−40.1	normal	−0.144(0.011)	8.5×10^{-37}
15	2159p484	62.2	normal	−0.143(0.009)	3.1×10^{-56}

Construction. Standardize `fracflux` and `rchi2` against linear-in-|BGAL| baselines and robust spreads ($\pm 1\sigma$ from the 16th/84th percentiles) fit across the other fields in the leave-one-out sample:

$$\zeta_{\rm ff} = \frac{\texttt{fracflux} - \mu_{\rm ff}(|\mathrm{BGAL}|)}{\sigma_{\rm ff}(|\mathrm{BGAL}|)}\,, \qquad \zeta_{\rm rc} = \frac{\texttt{rchi2} - \mu_{\rm rc}(|\mathrm{BGAL}|)}{\sigma_{\rm rc}(|\mathrm{BGAL}|)}\,.$$

`rchi2`'s coefficient never changes sign across any of the fifteen fields (Appendix H, Table H1), so $-\zeta_{\rm rc}$ is used the same way everywhere. `fracflux`'s contribution is sign-switched at $|\mathrm{BGAL}^\star| = a/|b| = 31.3°$, the zero-crossing of Eq. 9 itself:

$$Q = -\zeta_{\rm rc} + \zeta_{\rm ff} \quad (|\mathrm{BGAL}| \geq |\mathrm{BGAL}^\star|)\,, \qquad Q = -\zeta_{\rm rc} - \zeta_{\rm ff} \quad (|\mathrm{BGAL}| < |\mathrm{BGAL}^\star|)\,.$$

Higher Q means cleaner relative to other sources at the same latitude.

Validation. For each field, $\mu_{\rm ff}, \sigma_{\rm ff}, \mu_{\rm rc}, \sigma_{\rm rc}$ and the hinge $|\mathrm{BGAL}^\star|$ were refit from the other fields only (a leave-one-out refit of Eq. 9 itself, mirroring Section 7.3's own check), Q was computed for every real per-source record in the held-out field, and a univariate cluster-robust logistic regression of the real outlier flag on Q alone was fit – the same sandwich machinery as every other result in this paper, applied to Q in place of raw `fracflux`. (The first version of this appendix, and version 0.1.0 of the reference implementation, fitted these baselines against signed rather than absolute latitude; see Defect 6 in Appendix L. All numbers in Appendices O–Q are for the construction as described here.)

All fifteen leave-one-out fields give Q the correct sign; all but one are significant at $p < 0.01$, the exception (Field 11) at $p = 0.037$. None is significantly wrong-signed – against raw `fracflux`'s four significantly wrong-signed fields and four null fields in the `fracflux`-only fits of Table 2 (seven fields are significant with the expected sign there). This is a held-out-calibration result, not an in-sample fit: it uses the full $\sim 5.68 \times 10^6$-record dataset underlying Table 1, and every number in Table O1 is reproducible from the saved per-field `.npy` arrays with no access to a field's own data in its own calibration. It does not protect the choice of the score's form (which covariates enter, and the sign switch), which was made after seeing the coefficients of all fifteen fields: it checks the calibration constants and the regime boundary, not the design of Q.

What this does not show. Q is a relative, standardized score, not a calibrated contamination probability, and it has been tested on exactly the fifteen fields, one survey, one magnitude slice, and one outlier definition used throughout this paper, with a regime boundary (31.3°) that is specific to the executed model (Appendix R) – not shown to generalize beyond that regime, and not a substitute for the independent, fresh-field pre-registration Section 11 specifies. A reference implementation (`catwise-cleaner`, Python, MIT-licensed, with this exact leave-one-out validation reproducible from its own bundled script) is released alongside this paper's code (Data Availability, below); it is not, as of this writing, published to any package index, and no claim to the contrary should be inferred. Appendix P applies this same deployed score to a real, large-scale catalog-cleaning comparison and reports an honest limitation of the fixed $Q \geq 0$ threshold used above.

APPENDIX P: A REAL CATALOG-CLEANING IMPACT CHECK USING THIS PAPER'S OWN DATA

The project's original ambition for this check was a before/after catalog-cleaning benchmark against specific, already-published extreme-proper-motion and hypervelocity-object discoveries – several of which are the very discoveries motivating this paper (Section 1). Two obstacles made that version impossible to execute honestly rather than merely inconvenient. First, the computing environment used for this analysis has no network path to any of the relevant bulk catalog services (the unTimely/CatWISE host at NERSC, the Gaia archive, VizieR, IRSA, or the VISTA Science Archive); this was confirmed directly, not assumed, via the environment's own connection-policy diagnostics, which log an explicit policy denial for each host. Second, and independently, a check of the text of the discovery papers themselves – including Burgasser et al. (2024)'s hypervelocity L subdwarf already cited in Section 1, and a dedicated multi-epoch-WISE ultracool-dwarf search that explicitly imposes a `fracflux`> 0.5

Table P1. Real catalog-cleaning impact, pooled by the normal/reversed regime of Eq. 9 (all 5,683,222 real epoch-level records, Table 1). "Miss-bad" = fraction of this paper's own real outlier-flagged sources a rule's keep decision would still retain; "disc.-good" = fraction of real non-outlier sources it would needlessly discard. p: paired McNemar test (McNemar 1947) on the two rules' disagreements, for the bad/good comparison respectively.

Regime	N	raw miss-bad	Q miss-bad	raw disc.-good	Q disc.-good	p (bad / good)
Normal (9 fields)	3,253,480	84.40%	25.84%	5.71%	55.17%	$< 10^{-300}$ / $< 10^{-300}$
Reversed (6 fields)	2,429,742	39.75%	43.42%	47.71%	49.27%	8.3×10^{-19} / 1.6×10^{-177}

cut on unTimely data (Karpov et al. 2025) – found that neither tabulates the per-object `fracflux/rchi2/qf` values a reader would need to recompute Q for their specific objects; those columns are retained only in the underlying bulk catalog files, not in the papers themselves. Inventing plausible-looking numbers for named real objects to fill that gap would be indistinguishable from fabrication, so that version of this check is not reported.

What follows instead uses only data already analysed elsewhere in this paper: the deployed Q score (Appendix O; identical calibration constants to the released `catwise-cleaner` package) and Karpov et al. (2025)'s own published `fracflux`> 0.5 threshold are each applied, as a binary keep/discard rule, to every one of the 5,683,222 real epoch-level records underlying Table 1, and compared against this paper's own real, already-computed ground truth: the frozen pipeline's leverage-corrected, cluster-robust $|z| > 3$ outlier flag. Because both rules are evaluated on the identical set of real sources, the correct comparison is a paired one; we use McNemar's test (McNemar 1947) throughout. This is a real, traceable impact comparison – not a new validation claim beyond Appendix O, since it uses that appendix's full-sample (non-leave-one-out) deployed calibration rather than refitting anything new.

Pooled result. Of the 86,807 real sources this paper's own test flags as outliers, raw `fracflux`> 0.5 alone would still retain 64.98% of them (56,406); $Q \geq 0$ retains 33.48% (29,066) (40,232 sources that Q alone catches and raw `fracflux` does not, against 12,892 the reverse way; McNemar $p < 10^{-300}$). Of the 5,596,415 real sources *not* flagged as outliers, raw `fracflux` discards 23.66% and Q discards 52.64% (McNemar $p < 10^{-300}$). The two rules therefore sit at very different operating points: $Q \geq 0$ keeps only sources at least as clean as the latitude-typical source, so it discards about half of all sources by construction (52.9% of all records), and its higher catch rate is paid for with a much higher discard rate. Pooled, this comparison shows what each fixed cutoff costs; it does not show that Q is the better score, which is a threshold-free question (Appendix Q). **The pooled numbers average over different regimes.** Table P1 breaks the same comparison out by the normal/reversed regimes of Eq. 9. In the nine normal-regime fields Q catches every real outlier that raw `fracflux` catches (no cases the other way) plus 28,724 more, at a good-source discard rate of 55.17% against raw's 5.71%. In the six reversed-regime fields the two cutoffs operate at nearly the same point (good-source discard 49.27% for Q, 47.71% for raw `fracflux`) and, pooled over the regime, raw retains slightly fewer real outliers (39.75% vs. Q's 43.42%). That pooled figure hides very different behaviour field by field: in Fields 3 and 8, where raw `fracflux` ranks backwards (Appendix Q), the cut `fracflux`> 0.5 discards 90.2% and 89.7% of the good sources, so its low miss rate there (11.1%, 15.8%) is mostly the result of discarding almost everything; in Field 5 it discards only 9.7% and retains 86.9% of the real outliers. In the normal regime it is $Q \geq 0$ that is the aggressive rule, discarding more than half of the good sources. None of this contradicts Q's correctly-signed regression coefficients (Appendix O): a significant coefficient is a statement about the whole continuous relationship, while a cutoff at $Q = 0$ is a statement about one operating point that has not been tuned to any target false-positive rate (as `catwise_cleaner.scoring.clean_mask`'s docstring says). The honest reading is that neither a global cutoff on Q nor one on raw `fracflux` is uniformly appropriate across the sky, and that a comparison at equal discard rates, or a threshold-free one, is needed (Appendix Q).

An illustrative case. The record with the largest `rchi2` in Field 3 (tile `1439m485`, $|\mathrm{BGAL}| = 2.7°$; chosen by that criterion alone, not by any performance measure) has `fracflux`$= 0.5003$, just above the literature threshold, so raw `fracflux` alone would keep it – but its `rchi2`$= 40.60$ is the largest in that entire field (field median 1.13), driving $Q = -89.0$, far below the $Q \geq 0$ cutoff. The record is also flagged by this paper's own outlier test ($|z| = 3.65$): a case that a cut on `fracflux` alone cannot see. **What this does and doesn't show.** This is a real, fully reproducible impact comparison on data this paper already collected and already validated Q on (Appendix O) – it is not a test on unseen fields (that remains Section 11's job) and not a cross-survey generalization (Gaia DR3, Pan-STARRS1, ALLWISE, and VIRAC cross-matches were examined and are not available in the computing environment used here either). It also does not show $Q \geq 0$ is the right operating threshold for every use case; Table P1 is direct evidence that it is not, and that whoever deploys `catwise-cleaner` on a new sample should pick and validate their own threshold, exactly as its documentation already recommends. Appendix Q asks, in a threshold-free way, whether that shortfall is a calibration problem or a deeper ranking problem.

APPENDIX Q: IS Q'S REVERSED-REGIME SHORTFALL A THRESHOLD PROBLEM OR A RANKING PROBLEM? AN AUC CHECK

Appendix P compared two fixed cutoffs ($Q \geq 0$ and `fracflux`> 0.5) that discard very different fractions of good sources, so its miss rates are not like-for-like. A fixed threshold also conflates two separate questions: whether a score *ranks* sources correctly (independent of any cutoff), and whether a *specific cutoff* on that score is well calibrated. We separate them here using the area under the ROC curve (AUC), computed per field via the Mann–Whitney U relation (Mann & Whitney 1947): AUC = Pr(score of a random real outlier > score of a random real non-outlier), with AUC = 0.5 meaning no discriminative

Table Q1. Per-field AUC (Mann–Whitney U relation) for raw `fracflux` and Q as rankers of this paper's own real outlier flag, both using the identical real per-source data as every other table. diff = $\mathrm{AUC}_Q - \mathrm{AUC}_{\mathrm{raw}}$.

Field	Tile	BGAL	Regime	n	n_{bad}	$\mathrm{AUC}_{\mathrm{raw}}$	AUC_Q	diff
1	0000m016	−61.6	normal	255200	1200	0.548	0.618	+0.070
2	0720m591	−38.7	normal	412506	12390	0.641	0.676	+0.035
3	1439m485	2.7	reversed	366448	11794	0.459	0.607	+0.148
4	0000m228	−77.5	normal	313518	1890	0.560	0.643	+0.083
5	0719p090	−22.3	reversed	377148	2406	0.544	0.538	−0.006
6	1439p090	40.6	normal	312810	1845	0.541	0.610	+0.069
7	2159m091	47.4	normal	145126	808	0.545	0.599	+0.054
8	2879m091	−8.7	reversed	82884	876	0.429	0.608	+0.179
9	0719p484	2.2	reversed	475794	6026	0.497	0.585	+0.088
10	1439p484	46.9	normal	429104	4307	0.587	0.653	+0.066
11	2159m485	11.6	reversed	588702	11228	0.493	0.583	+0.090
12	2879m485	−23.2	reversed	538766	5433	0.536	0.543	+0.007
13	0000m515	−63.8	normal	426474	4810	0.603	0.672	+0.069
14	0719m485	−40.1	normal	505624	12914	0.661	0.712	+0.051
15	2159p484	62.2	normal	453118	8880	0.645	0.689	+0.044

power and AUC < 0.5 meaning the score is, on average, backwards. Scores are oriented so that higher means more suspect: $-$`fracflux` for the raw rule, $-Q$ for the deployed score (both using the exact same real per-source data and the exact same real outlier flag as every other table in this paper). The deployed Q's calibration constants were fitted to all fifteen fields, so, unlike Table O1, this audit is in-sample with respect to that calibration; within one field the baselines are constants, so they affect the ranking only through the relative weight of the two covariates and through the regime switch. Table Q1 gives the result for all fifteen fields. Q has a higher AUC than raw `fracflux` in 14 of 15 fields, including five of the six reversed-regime fields; in four of those (Fields 3, 8, 9, and 11) raw `fracflux`'s own AUC is below 0.5 ($0.459, 0.429, 0.497, 0.493$), a threshold-free, rank-based reappearance of this paper's sign-reversal finding. Treating each field as the unit of replication – as every other cross-field test in this paper does – a paired comparison across all fifteen fields is significant (mean AUC $0.553 \to 0.622$; paired $t = 5.72$, $p = 5.3 \times 10^{-5}$; Wilcoxon signed-rank $p = 1.2 \times 10^{-4}$). Within the nine normal-regime fields alone the advantage is universal (9/9 fields, mean AUC $0.592 \to 0.652$; $t = 11.95$, $p = 2.2 \times 10^{-6}$; Wilcoxon $p = 0.0039$, the smallest attainable two-sided exact value at $n = 9$). Within the six reversed-regime fields the advantage is smaller, present in 5 of 6 fields (the exception, Field 5, is an effective tie: $\mathrm{AUC}_{\mathrm{raw}} = 0.544$ vs. $\mathrm{AUC}_Q = 0.538$), and reaches significance on the paired t-test ($t = 2.79$, $p = 0.038$) but not on the exact Wilcoxon sign test at this small sample size ($p = 0.062$) – reported as it is rather than rounded to whichever side looks better. What this implies for Appendix P: Q ranks real outliers better than raw `fracflux` in 14 of the 15 fields, though by a modest margin (mean AUC $0.553 \to 0.622$; an AUC of 0.62 is far from a clean separation), including most reversed-regime fields, where raw `fracflux` ranks backwards or at chance. The fixed-cutoff comparison of Appendix P is best read as showing that neither global cutoff is well calibrated across the sky, not as evidence that Q carries less information than raw `fracflux` in crowded fields. It does not fix that calibration problem – doing so honestly would require choosing and validating a regime- or latitude-specific operating threshold on data not used to build it, which is exactly the fresh out-of-sample test Section 11 already calls for and this paper does not have new fields to perform.

APPENDIX R: SPECIFICATION SENSITIVITY OF THE HEADLINE RESULT

This appendix asks how much of the result of Section 7.3 depends on which other terms are in the model. The executed model (Eq. 8) is compared with seven alternatives, fitted to the same saved per-source arrays with the same source-clustered inference. Knots of the $W1$ and K splines are always fixed from the full primary sample, as the protocol requires, including for the $K \geq 10$ subsample; rank-transformed covariates are ranks within each field divided by the field's record count. For each specification the fifteen field coefficients are regressed on |BGAL| exactly as in Eq. 9 (inverse-variance weights; standard errors scaled by $\sqrt{Q_{\mathrm{res}}/13}$). Every row is regenerated by `reproducibility/specification_sensitivity.py` from the saved arrays; the first row reproduces Table 1 and Eq. 9.

What is robust. In all eight specifications the field coefficients rise toward the plane: the slope against |BGAL| is negative in every row (-0.0289 to -0.0450 per degree), significant at $p < 10^{-3}$ after residual scaling in every row, and explains 58–77% of the between-field heterogeneity. Fields 3 and 8 ($|\mathrm{BGAL}| = 2.7°$ and $8.7°$) have significantly positive coefficients in all eight.

What is not. The number of significantly positive fields ranges from 2 to 5, and the latitude at which the fitted line crosses

Table R1. Specification sensitivity, all fifteen fields. Counts use the $p < 0.01$ classes of Section 4.9. "Zero crossing" is the latitude at which the fitted line of coefficient against |BGAL| crosses zero; "Het." is the share of between-field heterogeneity (Q) explained by that line.

Specification	correct	reversed	null	reversed fields	slope	zero crossing (°)	Het. (%)
Executed model (Eq. 8)	5	5	5	3,8,9,11,12	−0.0370	31.3	59.2
Same, without $W1$ and K splines	11	3	1	3,8,9	−0.0448	14.7	58.3
`fracflux` alone, no controls	11	2	2	3,8	−0.0450	5.6	65.3
`fracflux` + $W1$ spline	11	2	2	3,8	−0.0404	0.2	62.7
`fracflux` + K spline	11	3	1	3,8,11	−0.0367	13.0	77.5
`fracflux` + $W1$, K splines (Table 2)	7	4	4	3,8,9,11	−0.0351	22.0	68.0
Executed model, rank-transformed covariates	6	5	4	3,8,9,11,12	−0.0289	28.2	69.3
Executed model, $K \geq 10$ only	6	5	4	3,8,9,11,12	−0.0334	30.6	61.1

Table S1. Per-field $\hat{\beta}_1$(se) of the `fracflux` coefficient in the executed model (Table 1) and after adding the terms of the frozen protocol that it omitted: a natural spline in the coverage count `nm`, one in $|\beta_{\rm ecl}|$, or both. Last column: change from the executed model to the full-protocol form, in units of the executed model's standard error. Script: `reproducibility/fields_omitted_terms.py`.

Field	|BGAL|	executed	$+\,s$(`nm`)	$+\,h(\lvert\beta_{\rm ecl}\rvert)$	both	Δ/se
1	61.6	−0.067(0.198)	−0.059(0.199)	−0.055(0.198)	−0.048(0.198)	0.10
2	38.7	−1.032(0.083)	−1.048(0.081)	−1.026(0.083)	−1.036(0.081)	−0.05
3	2.7	2.187(0.108)	2.190(0.107)	2.150(0.108)	2.153(0.107)	−0.32
4	77.5	−0.442(0.174)	−0.423(0.174)	−0.438(0.174)	−0.420(0.175)	0.12
5	22.3	0.099(0.123)	0.117(0.124)	0.106(0.123)	0.123(0.125)	0.20
6	40.6	0.137(0.166)	0.143(0.167)	0.138(0.166)	0.145(0.167)	0.05
7	47.4	−0.132(0.225)	−0.128(0.225)	−0.151(0.223)	−0.147(0.223)	−0.07
8	8.7	2.421(0.348)	2.801(0.348)	2.417(0.347)	2.793(0.347)	1.07
9	2.2	1.070(0.092)	1.109(0.093)	1.069(0.092)	1.108(0.093)	0.41
10	46.9	−0.273(0.103)	−0.229(0.104)	−0.275(0.103)	−0.234(0.104)	0.38
11	11.6	0.839(0.176)	0.871(0.169)	0.821(0.176)	0.854(0.169)	0.09
12	23.2	0.386(0.084)	0.387(0.084)	0.386(0.084)	0.387(0.084)	0.01
13	63.8	−0.700(0.107)	−0.687(0.107)	−0.696(0.107)	−0.681(0.107)	0.18
14	40.1	−1.259(0.071)	−1.256(0.071)	−1.266(0.071)	−1.263(0.071)	−0.06
15	62.2	−1.011(0.082)	−0.990(0.082)	−1.010(0.082)	−0.989(0.082)	0.28

zero from 0.2° to 31.3°; the executed model, which was frozen before the analysis, sits at or near the top of both ranges. Removing the $W1$ and K splines, or all controls, moves the zero crossing to 14.7° and 5.6°. We therefore treat the monotonic trend as the finding, and the sign reversal as established near the plane only; the value 31.3° used to define the "normal" and "reversed" regimes in Appendices O–Q belongs to the executed model and is not a property of the sky. This does not affect the leave-one-field-out validation of Q, which uses the same outlier flag, but it means Q's switching latitude is a model-conditional choice.

What this does not do. It does not choose among specifications, and it does not run the sensitivity analyses that the protocol listed but we did not (tertile-binned covariates, a catalog-cross-match H_6 screen, neighbour-separation thresholds; Section 6).

APPENDIX S: THE FULL FROZEN-PROTOCOL MODEL, ALL FIFTEEN FIELDS

The frozen protocol (PREREGISTRATION Section 4) specifies, besides the terms of Eq. 8, a natural spline in the coverage count `nm` of each epoch-level record and one in $|\beta_{\rm ecl}|$; the per-field model executed for Table 1 omits both (Section 6, deviation (i)). With the raw catalogs of all fifteen fields re-processed (Appendix G), the two terms can be added back, and every field was refit with the pipeline's own logistic and source-clustered sandwich code (`reproducibility/fields_omitted_terms.py`). `nm` is the per-epoch catalog column of that name (4–77 over all records); $|\beta_{\rm ecl}|$ is the absolute ecliptic latitude of the track's first-epoch position, which varies by at most 2.2° within a tile. The per-source arrays rebuilt for this check are identical to the stored ones in all fifteen fields, the executed-model refit reproduces the saved coefficients and standard errors exactly (to the last stored digit), and neither new spline is degenerate (its knots are distinct) in any field.

Table S1 gives the result. Adding both terms changes $\hat{\beta}_1$ by at most 0.37 in absolute value (1.1 standard errors; Field 8) and by a median of 0.019; it changes the sign of $\hat{\beta}_1$ in no field. Almost all of the movement comes from s(`nm`) (largest change 0.38, Field 8); $h(|\beta_{\rm ecl}|)$ alone moves no coefficient by more than 0.04 (0.3 standard errors). One field crosses the $p < 0.01$ line: Field 10 ($p = 0.0080$ in Table 1, 0.0241 with both terms; *correct* would become *null*); its sign is unchanged.

The meta-regression of Eq. 9 (inverse-variance weights, standard errors scaled by $\sqrt{Q_{\rm res}/13}$, as in Section 7.3) refit to the full-protocol coefficients gives slope −0.0371 per degree of |BGAL| (−0.0370 for the executed model), intercept 1.172 (1.158), $p = 8.9\times10^{-4}$ (7.9×10^{-4}) and a zero crossing at 31.5° (31.3°); the leave-one-out sign prediction is right in 14 of 15 fields under

all four models. The omission of the two terms therefore does not account for the Galactic-latitude dependence reported in this paper. This tests the specification only: the data, the outlier definition and the other deviations of Section 6 are unchanged.