Appendix: Supplementary analyses

Companion to the main manuscript; section and table references of the form "Section X" or "Table Y" point to the main manuscript.

A Supplementary analyses

A.1 Formal verification details

Table 1: Quantified property pack on the BEIS pipeline, grounded in real released numbers. Each block is a first-order property over the realised partition; UNSAT marks a violated safety property (the failure family in parentheses), SAT marks a property that holds or a deliberate witness. Quantified blocks run on Z3 4.15.4; Yices2 2.7.0 accepts only the quantifier-free companions. A hierarchical-reconstruction block (Q11 in the pack) is an illustrative latent property with no realised instance in the audited BEIS year (no MSOA there has exactly one suppressed child); it is vacuously safe and omitted from the tabulated verdicts.

Block	**Property (failure family)**	**Quant.**	**Verdict**	**Solvers**
Q01	within-window suppression $k \geq 5$	$\forall$	sat	Z3
Q02	cross-window differencing (*composition*)	$\forall\forall$	unsat	Z3
Q03	m-invariance signature (*composition*)	$\forall\forall$	unsat	Z3
Q04	order-of-operations (*transform-order*)	$\forall$	unsat	Z3
Q05	composed-risk monotonicity (lemma)	$\forall\forall$	sat	Z3
Q06	specification-validity budget $\exists!$ ($W^* = 2$)	$\exists\forall$	sat	Z3
Q07/Q08	gap-localiser determinism	$\exists!$	sat	Z3
Q09	structural $\not\Rightarrow$ operational (witness)	QF	sat	Z3, Yices2
Q10	pipeline non-commutativity (witness)	QF	sat	Z3, Yices2
Q12	policy admits unsafe release (free counts)	QF	sat	Z3, Yices2
Q13	remediation R^* closes the gap	QF	unsat	Z3, Yices2
Q14	suppress/hash non-commutativity lemma	QF	sat	Z3, Yices2
Q15	GoiEner: muni-aggregation admits sub-threshold ship	QF	sat	Z3, Yices2
Q16	GoiEner remediation (pin suppress-after-join)	QF	unsat	Z3, Yices2
Q17	MaStR: PLZ-aggregation admits a singleton tuple	QF	sat	Z3, Yices2
Q18	MaStR remediation (constrain full QI tuple)	QF	unsat	Z3, Yices2

Formal-layer details: minimal cores, repair synthesis, and the GoiEner and MaStR refutations. Two properties make the localisation rigorous rather than a string lookup. First, the family-localising cores are *minimal*: a deletion-based minimal-unsatisfiable-subset check confirms each core ($Q02$–$Q04$, $Q13$) is minimal, so a family label sits in the core only when it is *necessary* for the contradiction, which is what makes the family verdict sound. Second, the repair is *derived*, not asserted: as Z3's interpolation interface is unavailable in current builds, its Horn/Spacer fixed-point engine synthesises a sufficient repair predicate R such that the documented policy together with R entails margin-safety (it derives the equal-counts constraint). The window-budget operator R^* (equal counts, or a gap of at least five) is the minimal correction set of the *remediated* instance $Q13$ (deleting it re-satisfies the unsafe witness $Q12$), not of the bare policy, which is satisfiable; R^* is one sufficient fix, not the unique one (a consistent cross-release perturbation scheme is an alternative the layer does not model, which is why the publisher-facing remediation offers either a budget or a consistent mechanism). The order-of-operations family is stated as a reusable operator non-commutativity lemma (block $Q14$: suppression and the hash join do not commute). The same data-independent refutation applies to the other two non-trivial divergences. For GoiEner, the documented municipality-aggregation rule with its CUPS-hash join does not pin the operation order, so the policy admits a release whose shipped partition falls below the 50k threshold ($Q15$, SAT); pinning the order closes the gap ($Q16$, UNSAT). For MaStR, the PLZ-aggregation rule for sub-30 kW plants constrains only the postcode, not the published (PLZ, capacity, technology, year) tuple, so the policy admits a singleton class ($Q17$, SAT); adding the missing tuple constraint closes it ($Q18$, UNSAT). In each case the verdict refutes the documented specification over all admissible releases, and the stated repair is the solver's correction set; synthesising the budget and remediations directly, by optimisation- and maximum-satisfiability-modulo-theories, is the natural next step.

A.2 Reference tables and terminology

Table 2 gives the extended-metric pack and Table 3 the full per-release quasi-identifier provenance; the terminology list between them defines the vocabulary used throughout.

Table 2: Extended-metric pack per release (companion to Table 6): entropy and distinct ℓ-diversity, β-likeness, and 1-Wasserstein t-closeness. (α, k)-anonymity α and δ-presence are 1.00 for every release where defined (systemic singleton coverage) and are omitted. "n/a" indicates the metric is undefined for the release.

Release	$\boldsymbol{\ell}_{\text{ent}}$	$\boldsymbol{\ell}_{\text{dist}}$	$\boldsymbol{\beta}$	$\boldsymbol{t}$
acn-data	1.00	1	18.23	334.1
beis-subnational	1.00	1	9.01	3414.5
entsoe-demand	n/a	n/a	n/a	n/a
goiener	1.00	1	n/a§	n/a§
greend	1.00	1	7.00	3.5
ideal	1.00	1	254.00	81.9
low-carbon-london	1.00	1	6.00	8163.4
mastr	n/a	n/a	n/a	n/a
ocpp-ev	1.00	1	9.00	4.5
opsd-household	1.00	1	0.00	0.0
uk-dale	1.00	1	4.00	2.0

§ GoiEner's released table lacks a per-CUPS sensitive-attribute time series; β and t default to the sentinel 'unknown' and are reported as n/a rather than as numeric leaks.

Terminology.

Divergence. A counted mismatch between a release's stated privacy claim and its measured posture, adjudicated per claim type by the pre-registered rule of Table 1.

Claim falsification. The formal-side outcome in which the documented policy abstraction fails to entail the claimed guarantee, witnessed by an UNSAT result whose labelled core names the failing assumption.

Structural uniqueness. Row-level uniqueness on the publisher's declared quasi-identifier set (a record with equivalence-class size $k = 1$), a property of the released file alone.

Operational re-identification risk. The estimated probability that a structurally unique record is also unique in the underlying population, computed with the Rocher et al. [1] Poisson population-uniqueness estimator, a property of the file plus a population and an adversary.

Documentation finding. The infrastructure-subject carve-out: a release whose unique tuple has no natural-person data subject (e.g., `entsoe-demand`) yields a documentation finding rather than a divergence, even at row-level $k = 1$.

Amplification case. A re-identification exposure that materialises only after linkage of a release's quasi-identifiers against an external or publisher-owned auxiliary source, reported as an empirical-side amplifier rather than a formal-side divergence.

Table 3: Quasi-identifier provenance per release. *Source:* PQ = publisher-quoted from verbatim methodology text; AC = auditor-committed (inferred from data structure). Cardinality is the number of distinct values seen in the audited file (`n/a` where the column is continuous and binned at audit time, or where the audited sample is too small to estimate).

Release	QI column	Source	dtype	Cardinality / binning
acn-data	station_id	AC	string	54
acn-data	weekday_pattern	AC	categorical	7
acn-data	kwh_delivered_bucket	AC	numeric (bucketed)	10 quantile bins
beis-subnational	lsoa_code	PQ	string	51,147
beis-subnational	year	PQ	integer	15 (2010 to 2024)
beis-subnational	sector	AC	categorical	1 (domestic; audited slice, full release 2)
entsoe-demand	bidding_zone	PQ	string	1 (DE-LU sample)
entsoe-demand	timestamp	PQ	datetime	61,488
goiener	hashed_cups	PQ	string (hash)	30,261
goiener	province	PQ	categorical	n/a
goiener	tariff_type	AC	categorical	n/a
goiener	contract_class	AC	categorical	n/a
greend	country	AC	categorical	2 (IT, AT)
greend	plug_count	AC	integer	8
greend	peak_appliance	AC	categorical	8
ideal	number_of_occupants	AC	integer	5
ideal	dwelling_type	AC	categorical	n/a
ideal	region_uk	AC	categorical	n/a
low-carbon-london	acorn_group	AC	categorical	7
low-carbon-london	tou_flag	AC	categorical	2
low-carbon-london	weekday_peak_kwh	AC	numeric (bucketed)	10 quantile bins
mastr	plz	PQ	string	n/a (sample-frame)
mastr	capacity_kw	PQ	numeric (bucketed)	n/a
mastr	commissioning_year	AC	integer	n/a
mastr	technology	AC	categorical	n/a
ocpp-ev	charge_point_id	PQ	string	10
ocpp-ev	weekday_pattern	PQ	categorical	7
ocpp-ev	peak_energy_per_session	PQ	numeric (bucketed)	10 quantile bins
opsd-household	sub_meter_count	AC	integer	n/a (small cohort)
opsd-household	instrumented_circuits	AC	integer	n/a
opsd-household	weekday_pattern	AC	categorical	7
uk-dale	instrumented_circuits	AC	integer	5
uk-dale	weekday_pattern	AC	categorical	7
uk-dale	peak_load_kw	AC	numeric (bucketed)	10 quantile bins

A.3 Abstraction soundness

Table 4: Abstraction soundness: each SMT assertion is implied by the documented policy (over-approximation), so UNSAT is a genuine specification gap, not an encoding artefact. Every non-goal assertion is licensed by a clause in the publisher's verbatim-quote archive, which discharges the over-approximation obligation and doubles as evidence of faithful claim extraction.

Assertion label	Pipeline step	Licensing source	Direction
`POLICY_SUPPRESS_t`	per-cell < 5 suppression	DESNZ sub-national methodology booklet	over-approx (implied)
`POLICY_SUPPRESS_tp1`	same, next window	DESNZ booklet	over-approx (implied)
`COMPOSE_WINDOW_differencing_margin_safe`	negated cross-window safety goal	(the guarantee under test)	goal negation
`R_STAR_window_budget`	proposed cross-release budget operator	synthesised, not in policy	remediation candidate

A.4 Closed form for the specification-validity budget

Treat the decile-transition matrix of Figure 2(a) as a time-homogeneous Markov kernel and write s_i for the probability that a subject stays in decile i. A subject's bin set grows once it ever leaves its starting decile, so the composed violation rate after w windows is

$$\mathrm{minv}(w) = 1 - \sum_i \pi_i \, s_i^{w-1}. \tag{1}$$

With equal-mass global deciles this gives 0.347 and 0.548 at $w = 2, 3$, close to the measured 0.38 and 0.60 (the small under-prediction is the time-inhomogeneity the homogeneous chain omits). Solving $\mathrm{minv}(W) = \tau$ for a representative stay probability $\bar{s}$ gives

$$W^*(\tau) = \left\lfloor 1 + \frac{\ln(1-\tau)}{\ln \bar{s}} \right\rfloor, \tag{2}$$

reproducing $W^* = 2$ at $\tau = 0.5$ and $W^* = 1$ at $\tau = 0.25$. The absolute decay rate is binning-dependent, but the BEIS-faster-than-EIA ordering holds under pooled global deciles at $q \in \{5, 10, 20\}$ (ratios 3.8×, 2.5×, 1.7×), so we report the comparison as an ordinal, binning-robust observation. The second-eigenvalue mixing rate is not a usable proxy: BEIS mixes more slowly than EIA-861 yet decays faster, because most decile-to-decile movement is concentrated in the middle of the distribution, while the extreme deciles remain stable.

A.5 Adapter validation and empirical robustness

Cross-validation and positive control. *Cross-validation against Java ARX 3.9.1.* The pure-Python pyARX-equiv adapter is cross-checked against the upstream `libarx-3.9.1` Java library on *all eleven* releases (projecting each file to its quasi-identifier and sensitive-attribute columns before handing it to ARX). The two implementations agree exactly on k for every release (11 of 11, all $k = 1$) and on distinct-ℓ for the nine releases that carry a sensitive-attribute column in the audited file ($\ell_{\text{dist}} = 1$ throughout); ENTSO-E (an aggregate series with no sensitive attribute) and the audited MaStR sample (whose operator-flag column is not in the released file) have no sensitive attribute, so ℓ is undefined for both back-ends. Per-release (records-total, equivalence-classes) counts match across the two back-ends, including on the large releases: (639 300, 639 300) on beis-subnational, (305 053, 70 077) on goiener, (61 488, 61 488) on entsoe-demand, (9 294, 9 245) on mastr, and (750, 273) on acn-data. The t-closeness column is not directly comparable because pyARX-equiv uses 1-Wasserstein (earth-mover) distance while libarx-3.9.1 uses total-variation distance; the agreement on (k, ℓ_{dist}) is the load-bearing cross-validation signal. The Java driver and a one-command re-run script ship with the reproducibility artefact. *Synthetic positive control.* The adapter's metric implementation is also gated by a hand-constructed 100-row synthetic release with known $k = 5$, $\ell_{\text{dist}} = 3$, $\alpha = 0.6$, $\delta_{\text{presence}} = 0.2$, $t = 0$, and $\beta = 0$; the pure-Python pyARX-equiv reproduces every one of these values exactly. The control rules out the possibility that the adapter's $k = 1$ verdicts on the publisher releases are a coding artefact of the metric implementation.

Cross-release linkage matrix. As a robustness check that complements the privacy-utility frontier of Section 4.2, we compute the cross-release linkage matrix. For each ordered pair of releases (R_i, R_j) in the corpus, the linkage matrix records: the intersection of the publisher-declared QI columns $\text{QI}(R_i) \cap \text{QI}(R_j)$ (the join key, if any), the within-release k for each side on the join key, and the worst- side $k_{\text{under\_join}} = \min(k_i^{\text{join}}, k_j^{\text{join}})$ as a worst-side upper bound on the joined-projection k (joining can only reduce equivalence-class size, so $\min(k_i, k_j)$ bounds the joined k from above). Of $11 \times 10 = 110$ ordered pairs, 12 carry a non-empty join key: opsd-household and uk-dale share `instrumented_circuits` and `weekday_pattern`; opsd-household, ocpp-ev, and acn-data share `weekday_pattern`; uk-dale, ocpp-ev, and acn-data share `weekday_pattern`; and the directional duals of these. The worst joint-release linkage is the (opsd-household, uk-dale) pair at $k_{\text{under\_join}} = 1$ on the (`instrumented_circuits`, `weekday_pattern`) tuple. The other ten non-empty pairs land at $k_{\text{under\_join}} \in \{5, 10, 11\}$, dominated by the size-limited household and EV-charging cohorts. The `cross_release_linkage` family in Table 6's gap column stays at zero across the corpus because no within-corpus pair drives a release's k below the publisher's stated k-bound (most claims are pseudonymisation-only or aggregation-only, not k-bound claims); the within-corpus pairs that do bring $k_{\text{under\_join}}$ below 5 (the strictest stated bound) sit on releases whose own per-release k is already 1, so the linkage step adds no new divergence on top of the per-release row. The matrix is therefore a robustness check, not a primary result: it confirms that the audit's per-release verdicts are not artefacts of joining-blindness within the corpus. The `cross_release_linkage` family is elevated to two amplification cases under projection against external auxiliary sources in Section 4.3, which we report as an empirical-side amplification rather than a formal-side refutation because the property packs as released do not assert the external aux sources as labelled assertions.

QI-sensitivity sweep. A single-column QI sensitivity sweep over all eleven releases (33 measurement rows in the QI-sensitivity report) confirms that the four releases whose baseline $k = 1$ cannot be recovered by removing any one quasi-identifier are `opsd-household`, `mastr`, `greend`, and `goiener`; these are the four releases with row-uniqueness-driven $k = 1$ in the corpus. By contrast, seven of the eleven releases respond to a single-QI drop. The largest jump is BEIS sub-national: removing `lsoa_code` lifts k from 1 to 41,730 (a delta of +41,729): the audited BEIS slice is domestic-sector, so the residual quasi-identifier is effectively the year, and the 51,147 distinct LSOA codes (which appear across the 15 years, about 41,730 to 43,065 of them present in any one year) collapse onto the single year axis. Both `lsoa_code` and `year` are *publisher-quoted* QIs (the DESNZ methodology booklet names them; Table 3), so the two robust divergences (BEIS and GoiEner) rest on the publishers' own declared identifiers rather than on auditor-committed QI choices; the sensitivity sweep is what separates these from the auditor-committed-QI releases. Other one-column drops: `entsoe-demand`'s `timestamp` drop lifts k to 8784, `ideal`'s `number_of_occupants` drop lifts k to 11, `ocpp-ev`'s `charge_point_id` drop lifts k to 10, `acn-data`'s `station_` `id` drop lifts k to 6, `low-carbon-london`'s `weekday_peak_kwh` drop lifts k to 7, and `uk-dale`'s `instrumented_circuits` drop lifts k to 5. The seven responsive releases have a $k = 1$ verdict driven by exactly one quasi-identifier and are in principle remediable by coarsening that column.

β-likeness binning sweep. Re-running β-likeness with the qcut $q \in \{5, 10, 20\}$ on the three releases with a numeric sensitive attribute shows that the metric is stable on `low-carbon-london` (spread = 0; the weekly-consumption distribution is coarse enough that every q produces the same bin partition) but rises monotonically with q on `beis-subnational` (4.0 at $q = 5$ to 19.0 at $q = 20$) and on `acn-data` (18.2 to 374.0). The implication is that a single point-estimate β without a q-binning grid is brittle for sparsely-populated numeric channels; future audits should report $\beta_{q=5}$ and $\beta_{q=20}$ alongside the default $\beta_{q=10}$ whenever the sensitive attribute is numeric. The IDEAL $\beta = 254$ from Section 4.1 is excluded from this sweep because IDEAL's sensitive attribute is categorical (a survey-response code), so the qcut parameter has no effect: the reading is a categorical-frequency outcome, not a binning artefact.

References

[1] L. Rocher, J. M. Hendrickx, Y.-A. de Montjoye. *Estimating the success of re-identifications in incomplete datasets using generative models.* Nature Communications, 10: 3069, 2019.