Article

Not peer-reviewed version

Hierarchical Reconciliation of Fifty-One Years of Highway–Rail Grade Crossing Data with Fuzzy and AI Methods

Raj Bridgelall *

Posted Date: 11 March 2026

doi: 10.20944/preprints202603.0832.v1

Keywords: highway–rail grade crossings; railroad safety analytics; exposure normalization; crash frequency modeling; spatial risk assessment; data harmonization; inventory data quality; longitudinal infrastructure analysis; incident–inventory linkage; safety performance measurement

 doi:10.20944/preprints202603.0832.v1

Article

Hierarchical Reconciliation of Fifty-One Years of Highway–Rail Grade Crossing Data with Fuzzy and AI Methods

Raj Bridgelall

Department of Transportation and Supply Chain Management, College of Business, North Dakota State, University, P.O. Box 6050, Fargo, ND 58108-6050, USA; raj@bridgelall.com

Abstract

Highway–rail grade crossing (HRGC) safety research relies on federal incident and inventory datasets that span multiple decades. However, inconsistencies in geographic identifiers and incomplete reconstruction of crossing denominators can distort exposure-based rate metrics. This study develops, documents, and validates a reproducible nine-stage reconciliation pipeline applied to 51 years (1975–2025) of national HRGC incident data from the Federal Railroad Administration Form 57 and Form 71 datasets. The hierarchical pipeline integrated deterministic alignment and AI-assisted inference to produce an audited, geographically consistent dataset. The study formalizes four longitudinal county-level exposure metrics that quantify spatiotemporal risk. These metrics include accumulated incidents per million population (AIPM), accumulated incidents per crossing (AIPC), crossings per million population (CPM), and crossings per 100 square miles (CPHSM). All four metrics exhibited pronounced right-skewness: AIPM, CPM, and CPHSM approximated exponential forms, and AIPC approximated a log-normal form. Anderson–Darling tests detected statistically significant tail deviations in three metrics; CPM did not reject the exponential fit at conventional significance levels. Spatial analysis shows coherent regional concentration in incident rates in the Central Plains and lower Mississippi corridors. The national time series exhibits a late-1970s plateau, sustained exponential decline beginning around 1980, and stabilization but persistent incident rates after 2001. Population-normalized AIPM remained statistically indistinguishable between the reconciled and record-dropped datasets; however, crossing-based metrics changed materially when reconstructing denominators from the reconciled crossing universe. Median ratio comparisons confirmed that incident-only denominators introduced substantial measurement bias in local risk assessment. State-level rank reversals persisted even when omnibus distributional tests failed to reject equality. By formalizing multistage data cleaning and quantifying its analytical impact over an unprecedented longitudinal horizon, this study establishes denominator integrity and geographic reconciliation as prerequisites for valid HRGC exposure assessment and provides a replicable platform for future predictive modeling.

Keywords: highway–rail grade crossings; railroad safety analytics; exposure normalization; crash frequency modeling; spatial risk assessment; data harmonization; inventory data quality; longitudinal infrastructure analysis; incident–inventory linkage; safety performance measurement

1. Introduction

Highway–rail grade crossings (HRGCs) remain one of the most persistent interface risks in the U.S. surface transportation system. Federal Railroad Administration (FRA) datasets document more than five decades of collisions between rail equipment and highway users [1]. Substantial infrastructure investment and regulatory intervention occurred over that period, yet risks persist [2]. National summaries have consistently documented long-run reductions in incident frequency since the late 1970s, when annual totals exceeded 13,000 [3]. However, by the mid-2020s, approximately

 doi:10.20944/preprints202603.0832.v1

2,000 incidents continue to occur annually [4]. Residual incidents concentrate geographically and vary with infrastructure density and demographic scale.

Despite the availability of authoritative federal data, scholarly and applied studies typically emphasize incident modeling, risk factor identification, or intervention effectiveness. The available literature has devoted far less attention to the structural integrity of the underlying geographic identifiers that tie incidents to locations [5]. Data cleaning [6] and harmonization [7] consume the majority of analytical effort in practice, while model estimation occupies a secondary share. Published transportation safety studies seldom reflect this disproportion and rarely document these foundational processes in sufficient technical depth. Unexamined identifier misalignment and inconsistent denominator construction therefore distort exposure-based safety metrics.

Limited temporal scope constitutes a second research gap. Many published HRGC studies analyze relatively short time windows [8]. These studies focused on specific corridors or states or relied on cross-sectional slices of inventory data. Multi-decade longitudinal analyses do exist in federal reports and selected academic work [9]. However, they often aggregate trends without formally documenting the intermediate data engineering required to reconcile evolving identifiers, reporting formats, and infrastructure records. No prior study systematically documented a reproducible multistage reconciliation pipeline covering the full 50-year national HRGC record while reconstructing crossing denominators for county-level exposure metrics. The 51-year horizon from 1975 through 2025, combined with explicit auditing of each correction stage, represents a distinctive contribution.

This study addresses a narrowly defined measurement problem: whether multistage geographic reconciliation and denominator reconstruction materially alter exposure-based HRGC safety metrics. The research question was methodological rather than predictive: the study did not build a forecasting model but instead determined whether exposure-based metrics changed materially when identifiers were reconciled and when the crossing universe was reconstructed from both incident and inventory sources.

This study pursued three objectives:

1. Developing and documenting a reproducible, hierarchical data engineering pipeline that reconciled geographic identifiers across five decades of incident data while preserving auditability.
2. Reconstructing a defensible crossing denominator by integrating incident-referenced crossings with the federal inventory to form a unified crossing universe.
3. Quantifying the statistical and ranking impact of this reconciliation relative to a baseline record-dropping approach.

The contribution is methodological, empirical, and foundational.

Methodological contribution. The study formalized a nine-stage reconciliation hierarchy that combined deterministic FIP code alignment, fuzzy string matching, XID linkage, station-based inference, weighted multi-signal scoring, structured manual search, and controlled AI-assisted hypothesis testing. Each stage retained provenance labels. This design enabled transparent comparison between the baseline (GOOD) and reconciled (FIX) datasets. Explicit cleaning documentation at this level of detail remains uncommon in transportation safety research despite its central importance to analytic validity—a gap this study directly addresses.

Empirical contribution. The intermediate outputs constitute substantive findings. These outputs included distributional forms, spatial concentration, national temporal phases, state-level convergence, and denominator-sensitive metric divergence. The statistical comparison demonstrated that population-normalized incident rates remained stable under simple record-dropping. However, crossing-based exposure metrics changed materially when reconstructing denominators from the reconciled crossing universe. Hence, this study contributes to exposure measurement theory in infrastructure safety analytics. The results provide immediate insights for agencies that rely on exposure-based prioritization.

Policy relevance. The spatial and temporal patterns extracted from the cleaned dataset reveal coherent regional structures and a long-run exponential decline followed by a structural plateau.

 doi:10.20944/preprints202603.0832.v1

Even without extending into predictive modeling, these descriptive outputs inform strategic targeting of infrastructure upgrades, corridor diagnostics, and program evaluation. The rank-change analysis further showed that comparative state positioning shifted under rigorous reconciliation even when distributional tests indicated statistical similarity. Such sensitivity carries direct implications for resource allocation frameworks that rely on relative ordering.

Foundational contribution. By limiting scope to data cleaning and impact assessment, the study establishes a defensible analytic foundation for subsequent predictive, causal, or spatial econometric models. Clean, reconciled, and denominator-consistent county-level metrics are prerequisites for credible regression, machine learning, or risk-forecasting applications. This work, therefore, serves as a foundational platform for future modeling rather than a terminal study.

The scope is intentionally constrained. This paper does not attempt to estimate behavioral risk models, simulate intervention counterfactuals, or construct predictive algorithms. Such analyses would lack a defensible foundation without first establishing identifier integrity and denominator consistency across the full longitudinal record. Although data preparation often dominates analytical effort, it is seldom reported with technical rigor. Hence, a focused study on reconciliation and exposure reconstruction provides a strong and necessary contribution.

The remainder of this paper proceeds as follows. Section 2 reviews literature on large-scale transportation data cleaning and HRGC safety analytics. Section 3 describes the multistage reconciliation framework and denominator construction procedures. Section 4 presents the results, which include distributional modeling, spatial structure, temporal trends, and statistical impact assessment. Section 5 interprets the findings and discusses methodological and policy implications. Section 6 concludes by summarizing contributions and outlining future predictive research enabled by the cleaned dataset.

2. Literature Review

2.1. Large-Scale Safety Data Conditioning as an Analytic Determinant

Transportation safety analytics increasingly relies on administrative systems designed for operations, compliance, and reporting—not for research. These systems contain systematic linkage defects, missing values, and coding drift that directly distorts inference. HRGC research is especially vulnerable because prediction and screening models require cross-file joins between incident records and the FRA crossing inventory, making identifier consistency a prerequisite for valid inference [10]. Evidence on inventory quality shows that missing or inaccurate crossing attributes can change estimated relationships and degrade prediction accuracy [5]. Geographic reconciliation and denominator reconstruction therefore constitute first-order validity requirements for unbiased safety inference.

Outcome linkage research in road safety reinforced the same conclusion [11]. Police reports under-capture clinically relevant outcomes. Hence, linkage to EMS, trauma registries, or hospital admissions corrects outcome misclassification only when linkage bias is tested and documented [12]. Systematic syntheses show that linkage design choices and quality controls can dominate injury surveillance conclusions [13]. These findings generalize to HRGC analysis because exposure and severity inference depend on consistent identifiers and audited integration procedures.

2.2. HRGC Analytics for Prediction, Screening, and Decision Support

HRGC prediction studies have advanced from classical regression toward algorithmic screening and interpretable decision support. Early national and regional prediction frameworks estimated crash likelihood or frequency from inventory attributes and operational variables [9]. These studies emphasize that model accuracy depends on stable covariate encoding and reliable crossing identification, a dependency that practitioners often assume rather than verify. Modern classifiers report improved predictive performance relative to simpler baselines [14], but their feature sensitivity increases the cost of inconsistent coding and missing values [15].

 doi:10.20944/preprints202603.0832.v1

Recent journal studies extend prediction into deployment contexts. State-of-practice work used empirical methods [8], competing risk models [16], and machine learning [17] for safety enhancement at crossings. These studies assumed that the merged incident–inventory file was internally coherent. Other studies formalized alternative model structures and comparison frameworks for HRGC accident prediction [18]. Interpretable modeling emphasizes decision relevance by making risk factors legible to agencies [19]. This further heightens the need for traceable data conditioning. Regional benchmarking adds another layer by comparing geographic units under heterogeneity [20]. Denominator and location attribution consistency therefore becomes a prerequisite for meaningful geographic comparison. These approaches depend on consistent crossing counts, stable location coding, and defensible denominator construction; rankings shift when inventories are incomplete or denominators rely on incidents alone.

2.3. Severity and Consequence Analytics

HRGC severity studies show that risk mechanisms vary by crossing context, vehicle type, and roadway configuration. Private crossings differ from public crossings in operating controls, access patterns, and user expectations. This leads to different severity contributors and motivates stratified inference [21]. Heavy-vehicle involvement introduces maneuver constraints and kinetic differences that shift injury severity patterns relative to passenger-car cases [22]. Cross-category severity work for truck-trailer and passenger car crashes at railroad crossings reinforces that risk predictors are context dependent and interact with operational features in the crossing environment [23]. Configuration-focused work also narrows the design space by isolating specific roadway settings [24]. This enables cleaner inference at the expense of generality.

The literature also emphasized unobserved heterogeneity and spatial instability. Random-parameter approaches with heterogeneity in means and variances show that fixed-parameter severity results can conceal meaningful subgroup differences and can produce unstable marginal effects when applied across jurisdictions [25]. Fatal-crash studies highlight that the most severe outcomes occur under factor combinations that are not well represented by average conditions [26]. This increases the importance of correct attribution and denominator definitions in long-run screening.

Human factors research provides complementary mechanisms. Inattention is a persistent contributor at crossing [27]. Also, self-reported inattentive driving is systematically associated with crossing contexts. Pedestrian distraction at HRGCs introduces timing-based exposure risk during warning phases [28]. This becomes more salient as pedestrian flows grow near urban rail systems. These results indicate that behavioral mechanisms remain important even when engineering protections exist. This insight supports the observed persistence of residual incidents in mature safety regimes.

2.4. HRGC Research and Transferable Insights

HRGC studies outside the U.S. offer transferable insights about warning technologies, performance parameters, and measurement approaches. Technology adoption reviews argue that low-cost warning devices can expand coverage but require reliability and governance structures that manage fail-safe constraints, liability, and maintenance regimes [29]. Literature reviews synthesized parameters that influence HRGC performance and highlighted measurement gaps that can be traced to incomplete data capture and inconsistent reporting [30]. Macro-level studies linked system delay, traverses, and fatalities to HRGC risk [31]. They illustrated that performance metrics can be jointly determined by exposure and network operations.

Methodologically, several recent studies integrated new sensing and mobility data. Connected vehicles and GIS integration support behavioral measurement at crossings and creates richer exposure proxies [32]. However, they introduce additional linkage and cleaning demands due to multi-source fusion. Visibility measurement using LiDAR provides a remote inspection pathway for sightline auditing [33]. This supports infrastructure monitoring at scale while also creating new data-conditioning challenges in geospatial processing. Prediction work in emerging contexts extends

 doi:10.20944/preprints202603.0832.v1

crossing accident modeling into data-limited environments [34]. This reinforces the need for robust cleaning protocols when administrative data are sparse or inconsistent.

2.5. Positioning and Gap Addressed in the Present Study

Across HRGC and related research, two patterns recur. First, prediction, severity modeling, and prioritization have advanced rapidly [17]. Many studies now target operational decision support rather than only explanatory inference [19]. Second, most studies still treat data conditioning implicitly even though evidence shows that data quality and linkage can materially alter model estimates, predicted risk, and rankings [5]. In addition, the rise of multi-source fusion expands analytic capability while increasing the number of potential failure points in identifier alignment and denominator construction [33].

The present study addresses a narrow but high-value research agenda. Specifically, this is a cleaning-centered, audit-ready HRGC study that quantifies how reconciliation and denominator reconstruction alter exposure-based metrics, spatial patterns, temporal trends, and rankings over long horizons. That agenda complements existing modeling work by supplying a defensible analytic foundation for subsequent predictive and causal studies. It also produces intermediate descriptive outputs that are decision relevant.

3. Methodology

3.1. Data Sources and Analytical Framework

The analysis integrated two authoritative FRA datasets: the USDOT HRGC Incident File (Form 57) [1] and the HRGC Inventory File (Form 71) [35]. The incident file contains all collisions between railroad on-track equipment and highway users reported to the FRA since 1975. As of February 2026, the incident file comprised 250,090 records with 154 variables per incident. The inventory file contained 438,465 crossing records with 263 attributes describing location, operational characteristics, and physical features of crossings. Both datasets included counties outside the contiguous United States (CONUS); the pipeline subsequently removed them. Specifically, excluding Alaska, Hawaii, and non-state territories ensured spatial comparability across analytical units.

Although both datasets were nationally comprehensive, they were not designed for direct relational analysis. The incident file records event-level information, while the inventory file records crossing-level infrastructure characteristics. Key identifiers—crossing ID (XID), five-digit county FIPS code (FIP5), state name, county name, and nearest timetable station—contain inconsistencies, formatting variations, missing entries, and legacy coding artifacts inherited from decades of manual data entry and reporting system migrations. Moreover, crossings appearing in incident reports are not always present in the inventory extract, and vice versa. Therefore, direct aggregation without reconciliation can bias denominator-based safety metrics.

To address these structural limitations, the study developed a reproducible, code-based data engineering pipeline in Python. The pipeline performed multistage geographic harmonization, identifier canonicalization, crossing reconciliation, and metric construction prior to statistical analysis. The objective was not merely data cleaning but defensible reconstruction of rate denominators. This design ensured that incident counts and crossing counts were computed over a consistent and defensible spatial frame.

The cleaning pipeline proceeded through nine stages, illustrated in Figure 1. Formalizing each transformation step and retaining audit labels for all corrections established both traceability and reproducibility. This structure allowed subsequent statistical comparisons to isolate the methodological impact of rigorous reconciliation relative to baseline approaches that simply removed incomplete records.

 doi:10.20944/preprints202603.0832.v1

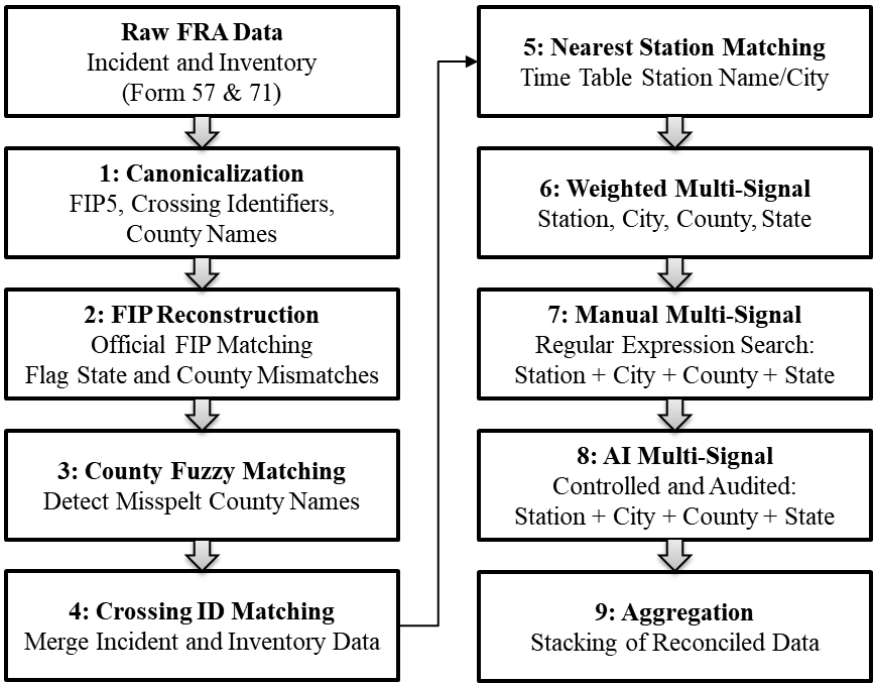

Figure 1. The Multistage data cleaning pipeline.

3.2. Multistage Geographic and Identifier Reconciliation

Structural, coding, and quality differences distinguish the incident and inventory datasets. A single-pass approach could not reliably resolve these discrepancies without either discarding substantial information or introducing geographic distortion. Therefore, the methodology applied a staged reconciliation framework that progressively resolved inconsistencies while preserving maximum usable data.

The reconciliation process followed three principles. First, corrections used authoritative geographic standards such as FIPS codes and official county boundaries wherever deterministic resolution was possible. Second, the pipeline flagged ambiguous records and resolved them using structured logic rather than ad hoc manual edits. Third, all transformations remained auditable to support replication and sensitivity testing.

The pipeline evaluated each incident record for consistency across multiple geographic fields. These included state name, county name, city name, nearest timetable station, and crossing ID. The pipeline applied corrections only when sufficient corroborating evidence existed across multiple fields. When deterministic resolution was not possible, the pipeline invoked secondary methods—fuzzy string similarity and multi-signal inference—within a controlled hierarchy.

The staged approach limited error propagation. Early stages addressed formatting and structural inconsistencies. Intermediate stages resolved spelling variations and cross-field mismatches. Final stages addressed rare edge cases requiring contextual reasoning. The pipeline flagged unresolved records rather than forcibly reassigning them.

This structured hierarchy served two purposes. It maximized recovery of valid geographic information and enabled transparent comparison between the baseline cleaning strategy—dropping records with incomplete or inconsistent fields—and the proposed multistage reconciliation process. Distributional tests and ranking comparisons in later sections evaluated the impact of this methodological choice. The following subsections describe each reconciliation stage in detail, beginning with geographic normalization and canonicalization procedures. Let D_0 denote the raw incident dataset indexed by record i. Each stage k produced two disjoint subsets:

$$D_k^{good} = \{ i \in D_{k-1} \mid \text{correction accepted} \}$$
$$D_k^{bad} = D_{k-1} \backslash D_k^{good} \quad (1)$$

 doi:10.20944/preprints202603.0832.v1

The pipeline terminated when no further reliable corrections remained.

3.2.1. Stage 1: Geographic Normalization and Canonicalization

Stage 1 established a consistent geographic foundation for both datasets. The objective was to eliminate formatting variation and legacy coding artifacts before attempting any inferential correction. Stage 1 standardized all state names to uppercase and restricted the analysis to CONUS states. Stage 1 canonicalized county identifiers (FIP5) by retaining only numeric characters and left-padding values to five-digit SSCCC format. Stage 1 converted nonconforming entries to null rather than coercing them into arbitrary codes.

The pipeline normalized crossing identifiers and simplified the attribute name to XID. It removed trailing decimal artifacts from spreadsheet exports (e.g., ".0"), stripped whitespace, and converted identifiers to a consistent string format. This step prevented formatting inconsistencies from compromising relational joins between incident and inventory files. The procedure standardized county and city names to title case and stripped extraneous suffixes or whitespace variation. It removed common administrative suffixes (e.g., "County," "Parish," "Borough") where necessary to enable consistent matching against official FIPS crosswalks. The pipeline retained records with missing or structurally invalid geographic identifiers in subset D_1^{bad} for evaluation in the subsequent stage. Stage 1 placed the subset requiring no correction into D_1^{good} for later aggregation. This separation preserved the distinction between mechanical formatting corrections and substantive geographic inference. Isolating canonicalization from later reconciliation logic reduced false mismatches attributable to formatting differences alone. It ensured that computationally intensive downstream methods—fuzzy matching, multi-signal reasoning, and crossing reconciliation—operated on standardized fields rather than heterogeneous raw inputs.

3.2.2. Stage 2: Deterministic FIP Alignment

Stage 2 applied the most reliable correction mechanism available: deterministic reconstruction of the five-digit county FIPS code. When both the state code and county code fields were populated, a canonical FIP5 identifier was constructed as

$$\text{FIP5} = zfill_2(\text{StateCode}) \;||\; zfill_3(\text{CountyCode}) \tag{2}$$

where the $\text{zfill}_n(c)$ function padded code c with zeros to equal n digits. The state component was zero-padded to two digits and the county component to three digits before concatenation. This transformation preserved official SSCCC formatting and avoided ambiguity across states with overlapping county codes. Stage 2 joined the reconstructed FIP5 value to an official county reference table (OCRT) containing authoritative state–county–FIPS mappings [36]. The pipeline accepted records where numeric codes matched valid counties within the reported state.

To guard against silent numeric errors such as transposed digits or off-by-one county codes, the procedure compared the normalized incident county name against the official county name associated with the reconstructed FIP5 in the OCRT. Stage 2 flagged discrepancies and routed them to subsequent stages rather than overwriting them automatically. This control prevented error propagation from transposed digits or miskeyed county codes. The procedure entered records satisfying numeric alignment and name consistency into the accepted subset for this stage, D_2^{good} and the residual records in D_2^{bad} for further evaluation.

This stage exploited the deterministic structure of federal geographic coding standards, ensuring zero probabilistic misclassification. Stage 2 yielded high-confidence corrections at negligible computational cost by relying exclusively on exact numeric alignment. It also reduced the search space for later probabilistic or inferential procedures, thereby preserving traceability and minimizing unnecessary complexity in downstream stages.

 doi:10.20944/preprints202603.0832.v1

3.2.3. Stage 3: Fuzzy String Matching (County Name Resolution)

Stage 3 addressed records where the FIP5 code was structurally valid and matched an official county in the reference table but the reported county name string did not match the canonical name associated with that FIP5. In these cases, the numeric identifier was treated as authoritative, and the objective was limited to correcting the textual county name field rather than reassigning geographic codes.

Stage 3 normalized all reported county names by removing punctuation, collapsing whitespace, converting them to title case, and stripping administrative suffixes such as "County," "Parish," "Borough," "Census Area," and "Municipality." This stage also converted known special cases such as "District of Columbia" to "DC" prior to comparison to avoid systematic mismatches, retaining the label for audit traceability. The procedure then compared the normalized incident county name with the canonical county name corresponding to its FIP5 by using a token-based string similarity metric. Specifically, the method computed similarity by using the token-set ratio function RapidFuzzTokenSetRatio from the RapidFuzz library (version 3.14.3) [37].

Let $s(a, b) \in [0, 100]$ denote the similarity score between the normalized incident county string a and the canonical county string b. The token-set ratio compared unordered token sets rather than raw character sequences. This reduced sensitivity to word order and minor punctuation differences while penalizing missing or extraneous tokens. Stage 3 accepted a correction when s(a, b) ≥ τ, where τ = 90. Sensitivity analysis and manual auditing established τ = 90 as the threshold that produced a zero false-positive rate across the full candidate set. For instance, a threshold of τ = 85 increased the candidate pool but introduced a non-zero false-positive rate, driven by phonetic similarities across states (e.g., "Greene" vs. "Green"). Human-in-the-loop auditing and crosswalk verification against the OCRT confirmed that a threshold of τ = 90 consistently provided a zero false positive rate. The procedure classified records exceeding this threshold as spelling corrections, labeled them CNTY_SP, and stored them in D_3^{good} with the remaining records stored in a subset D_3^{bad} for processing in subsequent stages.

Examples of accepted corrections included resolving "De Kalb" to "DeKalb," "Green" to "Greene," "O Brien" to "O'Brien," and truncated strings such as "Prince Willi" to "Prince William." The risk of geographic misassignment was minimal because Stage 3 operated only when the FIP5 code was already validated. This step isolated orthographic inconsistencies from substantive geographic errors, preserving identifier integrity while improving textual consistency. The explicit similarity threshold ensured reproducibility.

3.2.4. Stage 4: Crossing ID (XID) Matching

Stage 4 used the unique XID to resolve geographic inconsistencies that remained after Stage 3. The XID served as a persistent infrastructure-level key linking incident records (Form 57) to the FRA Crossing Inventory (Form 71). When valid, it provided authoritative geographic attributes independent of textual county or city entries. Stage 4 used canonicalized XIDs from Stage 1 to ensure consistent string representation. For records unresolved in prior stages, the pipeline joined the cleaned XIDs to the deduplicated FRA inventory dataset, filtered to at-grade crossings. It collapsed duplicate inventory entries for the same XID by retaining the first valid occurrence to ensure one-to-one matching. For valid XID matches, the incident record inherited the canonical state name, county name, and FIP5 associated with that crossing in the inventory dataset. Stage 4 treated this reassignment as authoritative because the inventory served as the official registry of crossing attributes maintained by railroads and state agencies.

Records without a valid XID, with placeholder identifiers (e.g., "PRIVATE," "NOTASGN"), or whose XID was not present in the inventory dataset remained uncorrected in this stage and stored in a subset D_4^{bad} for subsequent inference stages. The pipeline identified that approximately 15% of incident-referenced XIDs were absent from the contemporaneous inventory extract. This reflected historical retirement, reporting lag, and inventory synchronization gaps, with the majority from Texas, Illinois, California, and Ohio in the pre-1990 era. Stage 4, therefore, resolved geographic

 doi:10.20944/preprints202603.0832.v1

ambiguity through infrastructure-level linkage rather than textual similarity. The procedure labeled corrected records "XID" and stored them in subset D_4^{good} for auditability.

3.2.5. Stage 5: Nearest Station Match

Stage 5 used the nearest timetable station field as an auxiliary geographic signal for records unresolved after deterministic FIP alignment, county spelling correction, and XID matching. The nearest timetable station field reflects operational rail geography—defined by railroad timekeeping and dispatching practice—rather than census-defined municipal boundaries. The nearest timetable station therefore provided a geographically independent locational signal.

The procedure normalized all station names using the same canonicalization steps applied to other textual fields. It then matched the incident station field against a deduplicated station reference table derived from the FRA inventory dataset. This table contained unique combinations of station name and state, each associated with a valid county FIP5. The procedure accepted a match only when both the normalized station name and the reported state were consistent with a unique station–state pairing in the reference table. This state constraint prevented cross-state ambiguity arising from station names that recur nationally. When Stage 5 identified a unique county FIP5 through this constrained match, the record received the corresponding canonical state and county identifiers. The procedure then labeled it as "STATION" and appended it to a D_5^{good} subset. The procedure entered the remaining records into the D_5^{bad} subset for subsequent inference procedures.

Stage 5 therefore used rail-operational geography to resolve administrative ambiguity. Unlike county or city names, timetable stations were defined within the rail network and frequently corresponded to specific infrastructure locations within a county. When uniquely identifiable within a state, the station field provided an independent corroborating signal that strengthened geographic reconciliation without introducing probabilistic inference.

3.2.6. Stage 6: Weighted Multi-Signal Matching

Stage 6 resolved residual records for which previous stages could not establish a unique county assignment. These records contained compound inconsistencies across multiple textual fields. Single-field comparison was simple but unreliable because it ignored corroborating information available across other fields. Stage 6 therefore applied a structured weighted multi-signal inference framework.

Lookup dataset: For each candidate c, the reference table utilized in Stage 1 provided the canonical state name, county name, city name, and county FIP5 code. The procedure precomputed normalized keys $(\tilde{s}_c, \tilde{k}_c, \tilde{v}_c, \tilde{n}_c)$ for each candidate record in the reference table to represent state, county, city, and station proxies, respectively. Normalization lowercased strings, removed punctuation, collapsed whitespace, and removed administrative suffixes such as "County," "Parish," "Borough," and "Census Area" from the county names. A special-case rule standardized known variants such as "St John the Baptist Parish" to "St John" before tokenization.

Field-wise similarity. For each incident record i, the algorithm formed normalized observed fields $(\tilde{s}_i, \tilde{k}_i, \tilde{v}_i, \tilde{n}_i)$ to represent the reported state, county, city, and station names. For any field pair (a, b), the algorithm computed token-set similarity $s(a,b) \in [0,100]$ from the token-based ratio of the RapidFuzz function. Token-set similarity compared unordered token sets rather than raw character sequences. This reduced sensitivity to word order, punctuation, and spacing, while still penalizing token omissions and additions.

Weighting and normalization. The algorithm began with a base weight vector

$$\mathbf{w}^{(0)} = \{w_{\text{state}}^{(0)} = 0.35, w_{\text{county}}^{(0)} = 0.25, w_{\text{city}}^{(0)} = 0.20, w_{\text{station}}^{(0)} = 0.20\}. \quad (3)$$

The algorithm renormalized weights because many erroneous incident records contained missing fields. Let $P_i \subseteq \{\text{state, county, city, station}\}$ denote the fields (signals) that are nonempty for record i. The renormalized weight for field j is

 doi:10.20944/preprints202603.0832.v1

$$w_{ij} = \frac{w_j^{(0)} \mathbb{C}(j \in P_i)}{\sum_{m \in P_i} w_m^{(0)}} \tag{4}$$

so that

$$\sum_j w_{ij} = 1. \tag{5}$$

If all fields were empty after normalization, the algorithm assigned a score of zero and routed the record to D_6^{bad}.

Candidate pool restriction. Exhaustive comparison against the full lookup table increased computational burden and the probability of spurious near-ties. Therefore, the algorithm restricted comparison to a candidate pool, selected using a hierarchical seed field ordered as station, city, county, then state. The algorithm extracted the top N candidates (e.g., $N = 60$) from the lookup table using fuzzy search on the seed field only. This step reduced the search space while preserving plausible matches and improved runtime by confining full multi-signal scoring to the limited candidate pool.

Composite scoring. For each candidate c in the pool, the algorithm computed field-level similarities between normalized incident and candidate values for state, county, city, and station using the token-set similarity function

$$s_{\text{state}} = s(\tilde{s}_i, \tilde{s}_c), s_{\text{county}} = s(\tilde{k}_i, \tilde{k}_c), s_{\text{city}} = s(\tilde{v}_i, \tilde{v}_c), s_{\text{station}} = s(\tilde{n}_i, \tilde{n}_c). \tag{6}$$

The similarity score is zero if the corresponding field was missing in record i. The algorithm then computed the weighted composite score:

$$S_{ic} = \sum_{j \in P_i} w_{ij} \; s_j(\tilde{x}_{ij}, \tilde{x}_{cj}), \; S_{ic} \in [0, 100]. \tag{7}$$

Let $c^{(1)}$ be the best-scoring candidate and $c^{(2)}$ be the second-best candidate. The algorithm computed the margin

$$\Delta_i = S_{ic^{(1)}} - S_{ic^{(2)}}. \tag{8}$$

The procedure accepted the correction by entering the row into D_6^{good} only if $S_{ic^{(1)}} \geq \tau$ and $\Delta_i \geq \delta$. As in Stage 3, the thresholds were set to τ = 90 and δ = 3 based on sensitivity analysis and manual auditing to eliminate false positives. The margin condition guarded against ambiguous matches if multiple candidates had very close scores under the available signals. Sensitivity tests varying τ between 85 and 95 produced no material change in state-level rankings or distributional conclusions. The accepted records inherited the candidate's canonical state name, county name, and county FIP code. The procedure stored the composite score and margin as diagnostics into D_6^{bad} and labeled the fix in D_6^{good} as MULTI_SG or ADJ STATE when the accepted match implied a state correction. Stage 6 was the most flexible yet rigorously controlled component of the reconciliation hierarchy. Integrating multiple corroborating signals under explicit weighting and margin constraints allowed Stage 6 to resolve complex edge cases while preserving auditability.

3.2.7. Stage 7: Manual Multi-Signal Inference

This stage fixed the small residual subset of records that remained unresolved in the previous stages. These records exhibited multiple missing or internally conflicting fields such that automated scoring could not produce a unique match above the predefined thresholds. Stage 7 applied a structured manual search protocol rather than ad hoc edits. First, it constructed a composite search string by concatenating the state, county, and city fields in the reference geographic source [36]. Based on observations of the fields in the unresolved records, the researcher applied a case-insensitive regular expression of the form

 doi:10.20944/preprints202603.0832.v1

$$(?i)(?=.*\text{keyword1})(?=.*\text{keyword2}) \quad (9)$$

to conduct a search on the composite string. The expression required the simultaneous presence of two or more keyword fragments. This allowed partial spellings and phonetic variants to surface plausible candidates.

The search intentionally used abbreviated keyword fragments to capture likely truncations or misspellings while limiting false positives. For example, fragments such as "cross" and "wis" consistently identified La Crosse, Wisconsin, while "wood" and "rankin" identified Flowood in Rankin County, Mississippi. The researcher then verified candidate matches using external geographic references and contextual knowledge of rail-served communities. Stage 7 accepted a correction only when multiple corroborating signals converged on a single plausible county-state pairing. If ambiguity remained, this phase retained the record as unresolved rather than forcibly reassigned.

Stage 7 prioritized interpretability over automation. Although manual, Stage 7 followed an explicit search protocol and required external verification before acceptance. Restricting manual inference to a small residual subset preserved scalability while resolving rare but consequential edge cases with high confidence. As in the previous stages, the procedure entered the fixed and residual records in their respective subsets and labeled the fix as "MSig_Search" for auditing.

3.2.8. Stage 8: Verified Generative AI Multi-Signal Inference

Stage 8 applied a verified generative AI (GAI) inference to resolve the final residual records that resisted all prior methods. These cases contained multiple missing fields, internally conflicting geographic signals, or names of localities that corresponded to unincorporated communities that were not easily resolved through structured lookup tables.

The pipeline employed GAI platforms from three providers: OpenAI [38], Google [39], and Anthropic [40]. Stage 8 treated GAI platforms as hypothesis generators only, requiring independent verification against authoritative FIPS crosswalks before accepting any correction. The prompt issued to the three platforms was:

> You are an expert in American geography. I'm cleaning a railroad incident dataset to recover missing or incorrect FIP codes. I will provide you, one by one, with a state name, a county name, a city name, and a city name of the nearest station reported, all likely erroneous. Where a field is empty, I will put a placeholder "missing". Please provide only the likely county, the FIPS code, and a single sentence reason for the choice, nothing more. (10)

For each record, Stage 8 compared candidate responses across all three platforms and accepted a correction only upon unanimous convergence followed by deterministic FIPS verification. The researcher manually validated every proposed correction against the authoritative geographic reference before acceptance. GAI output alone did not constitute acceptance; every proposed correction required independent verification against authoritative geographic identifiers.

This stage intentionally restricted the GAI queries to a very small subset of cases. GAI queries incurred computational cost and hallucination risk and were subject to token limits. Large-scale application would require extensive auditing and risked introducing systematic bias. The study, therefore, treated GAI as a bounded, last-resort inference mechanism within a transparent reconciliation hierarchy. By structuring GAI use within explicit validation constraints, Stage 8 preserved reproducibility while capturing geographically plausible matches beyond deterministic or algorithmic similarity methods. The researcher labeled all GAI-assisted corrections as "GAI" to enable downstream reliability comparison across fix types. Stage 8 resolved all remaining records; the pipeline thereby achieved 100% geographic assignment across incident records after CONUS filtering.

 doi:10.20944/preprints202603.0832.v1

3.2.9. Stage 9: Final Aggregation and Dataset Construction

After all reconciliation stages, the pipeline stacked accepted records from each stage into a unified cleaned (FIX) dataset. The pipeline appended records in hierarchical order of resolution to preserve traceability of the correction method applied at each stage. Each accepted record retained a label (Fix Type) identifying the stage responsible for its correction.

Stage 9 restricts the final dataset to CONUS states to ensure geographic consistency and preserve a spatially contiguous rail network. The pipeline removed records reassigned during correction to Alaska or Hawaii, yielding 9,422 reconciled records. The resulting FIX dataset reflected standardized geographic identifiers, reconciled crossing-level linkages, and explicit documentation of correction provenance. This structured aggregation enabled direct comparison with a baseline strategy that simply dropped inconsistent records (GOOD), thereby isolating the methodological contribution of the multistage reconciliation framework.

3.3. Crossing Count Accounting

County-level rate metrics required an accurate crossing denominator. In HRGC safety analysis, this denominator is the number of unique at-grade crossings within each county. Relying solely on incident-observed crossings systematically undercounted infrastructure exposure because incident records captured only crossings where incidents occurred—not the full at-grade crossing population. To address this limitation, the study constructed a unified crossing universe by reconciling the cleaned incident dataset (FIX) and the FRA Crossing Inventory. A filter restricted both datasets to at-grade crossings, consistent with FRA Form 71 classification codes, thereby excluding pedestrian crossings and grade separations from the crossing universe. The pipeline canonicalized XIDs and standardized FIP5 codes in both datasets before comparison. The pipeline removed duplicate entries within each dataset using unique (FIP5, XID) pairs. The comprehensive crossing set used for denominator construction was the union of both datasets with duplicate XIDs removed. The union-based approach eliminated two sources of denominator bias. First, it avoids undercounting counties where crossings exist but have not experienced recorded incidents. Second, it captures incident-reported crossings that are missing from the inventory extract due to historical retirement, reporting lag, or synchronization gaps.

Approximately 15% of unique crossings in the incident dataset were absent from the inventory extract. Conversely, crossings appearing in incident records represented only about one-quarter of the total crossings listed in the inventory. Neither dataset alone provides an unbiased estimate of the national at-grade crossing population. Crossings retired before 1975 and absent from both files remained structurally unobservable, introducing a left-censoring constraint on pre-1985 AIPC values. This limitation is discussed further in Section 5.

Crossing counts were essential for the construction of rate-based safety metrics. These included AIPC (accumulated incidents per crossing), CPM (crossings per million population), and CPHSM (crossings per 100 square miles). For the baseline GOOD dataset, crossing denominators used only crossings observed within retained incident records. This reflected the conventional strategy of dropping inconsistent data without infrastructure reconciliation.

This reconciliation step materially improves exposure measurement by supplying a defensible crossing denominator for all four county-level metrics. Without it, counties with many crossings but few incidents would appear artificially safe, while counties with concentrated incidents at a small subset of crossings would appear disproportionately risky. Nevertheless, by accounting for the available infrastructure universe, the study isolated the impact of data cleaning on both numerator (incident count) and denominator (crossing count) components of county-level safety metrics.

3.4. Annual Trend Modeling

The analysis aggregated annual incident counts to produce a time series y_t for $t = 1, \ldots, T$ spanning 1975–2025 (T = 51 years). The baseline linear trend model was

 doi:10.20944/preprints202603.0832.v1

$$y_t = \alpha + \beta t + \varepsilon_t \tag{11}$$

where α is the intercept, β is the annual rate of change, and ε_t is an error term. The Bai-Perron methodology identified structural breaks, allowing regression coefficients to change at unknown dates [41]. The piecewise linear specification is

$$y_t = \alpha_j + \beta_j t + \varepsilon_t, \qquad t = T_{j-1} + 1, \ldots, T_j, \qquad j = 1, \ldots, m+1 \tag{12}$$

where T_j are unknown breakpoints and each regime j has its own intercept α_j and slope β_j. The model selects the break dates to minimize the total residual sum of squares (RSS)

$$\mathrm{RSS}_m = \sum_{j=1}^{m+1} \sum_{t=T_{j-1}+1}^{T_j} \left(y_t - \alpha_j - \beta_j t\right)^2 \tag{13}$$

subject to a minimum segment length constraint to prevent overfitting.

To select the optimal number of breakpoints, the pipeline estimated models with m = 0, 1, and 2 breaks and compared them using the Bayesian information criterion (BIC):

$$\mathrm{BIC}_m = T \ln\left(\frac{\mathrm{RSS}_m}{T}\right) + k_m \ln(T) \tag{14}$$

where:

RSS_m is the residual sum of squares under $m$breaks,

k_m is the number of estimated parameters,

T is the number of annual observations.

For a piecewise linear model with m breaks

$$k_m = 2(m+1) \tag{15}$$

since each regime estimates one intercept and one slope. The preferred model is the one with the lowest BIC. Differences in BIC are interpreted as:

| ΔBIC | < 2: weak evidence

2–6: positive evidence

6–10: strong evidence

> 10: very strong evidence

A ΔBIC exceeding 10 when moving from one break to two provided very strong evidence that the improvement in model fit outweighed the parameter penalty. This criterion provided an objective basis for identifying a multi-phase structure in the annual HRGC incident trend.

3.5. Incident Rate Distribution Modeling

After geographic reconciliation and crossing count accounting, the study constructed four county-level safety metrics to evaluate exposure-adjusted risk patterns. These metrics incorporated incident counts as numerators and population or crossing counts as denominators, depending on the conceptual framing of exposure. The analysis applied formal distributional modeling prior to statistical comparison because rate metrics derived from sparse-count data typically exhibit heavy-tailed behavior.

Let I_c denote the total number of incidents in county c, P_c its population, and X_c the number of unique at-grade crossings derived from the reconciled crossing universe. The primary metrics were defined as follows:

Accumulated incidents per million population (AIPM)

Accumulated incidents per crossing (AIPC)

Crossings per million population (CPM)

Crossings per 100 square miles (CPHSM).

 doi:10.20944/preprints202603.0832.v1

Together, AIPM captured population exposure, AIPC captured crossing-level risk intensity, CPM captured infrastructure-to-population balance, and CPHSM captured spatial infrastructure concentration. The per-capita normalization established control for large population variations among counties that can bias interpretations. However, small rural counties with few crossings can exhibit high per-capita rates, while large metropolitan counties can dilute the number of incidents across large populations. Consequently, assuming normality would be inappropriate without empirical verification.

For each metric, the study evaluated four candidate theoretical distributions: the normal, log-normal, gamma, and exponential families. Maximum likelihood estimation fitted the distribution parameters. The study assessed goodness-of-fit using both the Kolmogorov-Smirnov (K-S) and Anderson-Darling (A-D) statistics to provide complementary diagnostics [42]. The K-S statistic measures the maximum absolute deviation between the empirical cumulative distribution function (ECDF), $F_n(x)$, and the fitted theoretical cumulative distribution function (CDF), $F(x)$ where

$$D = \sup_x |F_n(x) - F(x)|. \tag{16}$$

The null hypothesis stated that the observed data are drawn from the specified theoretical distribution. Failure to reject the null indicated that the fitted distribution cannot be statistically distinguished from the empirical distribution at the selected significance level. However, with large samples such as those exceeding 2,400 counties, the K-S test becomes highly sensitive and may reject even minor deviations from the theoretical form. Moreover, the K–S statistic is driven by the single largest vertical discrepancy between the ECDF and CDF and is relatively less sensitive in the distribution tails. To address these limitations, the procedure also computed the A-D statistic. The A-D statistic evaluates the integrated squared deviation between $F_n(x)$ and $F(x)$, weighted by the inverse variance of the theoretical distribution

$$A^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x)]^2}{F(x)[1 - F(x)]}\, dF(x). \tag{17}$$

This weighting increases sensitivity in the tails of the distribution, where $F(x)$ approaches 0 or 1. Tail behavior is substantively important in county-level safety metrics because extreme counties—those with unusually high incident rates or crossing densities—often drive risk interpretation and resource prioritization. Accordingly, the K-S test provides a general measure of overall distributional deviation, while the A-D test provides enhanced sensitivity to tail misfit. Using both statistics offers a more robust assessment of parametric suitability, particularly in large samples where purely mechanical rejection may otherwise occur.

To mitigate instability inherent in ratio-based safety metrics, the analysis imposed minimum exposure thresholds on the denominator variables prior to distribution fitting and hypothesis testing. Counties with very small populations, limited land area, or few crossings can produce disproportionately large rates from one or two incidents. These low values would reflect stochastic fluctuation rather than persistent structural risk. Such small-denominator effects introduce leverage distortion in maximum likelihood estimation and inflate tail-sensitive goodness-of-fit statistics. The selected thresholds therefore functioned as denominator-stability constraints that ensured each retained observation represented a minimally sufficient exposure base. A structured sensitivity assessment varying each cutoff by ±50% demonstrated that the principal distributional forms, parameter estimates, and FIX–GOOD comparisons remained substantively unchanged. This robustness indicates that the reported results are not artifacts of arbitrary trimming but reflects stable underlying patterns in county-level HRGC exposure and incident concentration.

Distribution modeling served two purposes. First, it characterized the structural form of county-level safety metrics after rigorous cleaning and denominator reconstruction. Second, it guided the appropriate selection of statistical tests for comparing the FIX and GOOD datasets. Where normality assumptions were violated, nonparametric tests ensured robustness. Modeling empirical

 doi:10.20944/preprints202603.0832.v1

distributions rather than imposing parametric assumptions ensured that downstream inference reflected the actual statistical behavior of the reconciled data.

3.6. Cleaning Impact Assessment

The cleaning impact assessment quantified how multistage reconciliation altered county-level safety metrics relative to a baseline strategy of retaining only internally consistent records. The comparison isolated the effects of structured geographic correction and comprehensive crossing count accounting.

3.6.1. Construction of Comparative Datasets

Let $\mathcal{D}_{FIX}$ denote the cleaned dataset (FIX) obtained after all reconciliation stages and crossing union construction. Let $\mathcal{D}_{GOOD}$ denote the baseline dataset that the procedure constructed by retaining only incident records with internally consistent geographic identifiers and computing crossing counts solely from those retained records. For each county c, define:

I_c^{FIX} and I_c^{GOOD} as the accumulated number of incidents from 1975 to 2025

P_c as their average recent 10-year population from the U.S. Census Bureau [43]

A_c is the land area of county c in square miles

X_c^{FIX} as their crossing count from the reconciled union set

X_c^{GOOD} as their crossing count derived only from retained GOOD incident records

The primary metrics computed for both datasets were:

$$\mathrm{AIPM}_c = \frac{I_c}{P_c} \times 10^6 \tag{18}$$

$$\mathrm{CPM}_c = \frac{X_c}{P_c} \times 10^6 \tag{19}$$

$$\mathrm{AIPC}_c = \frac{I_c}{X_c} \tag{20}$$

$$\mathrm{CPHSM}_c = \frac{X_c}{A_c} \times 100 \tag{21}$$

The study utilized a static long-run population baseline to normalize the 51-year cumulative incident totals. This method isolated the incident variance from temporal demographic shifts and established a fixed exposure index. By anchoring the denominator to the 2010–2020 census average, the analysis leveraged the period of highest data fidelity and provided a stable benchmark for contemporary risk assessment. This decadal average mitigated stochastic demographic fluctuations and historical census methodology shifts, ensuring a robust central tendency for the normalization constant. Because the FIX and GOOD subsets utilized an identical denominator, the choice of population vintage did not bias the relative state rankings or the outcomes of the distributional tests.

3.6.2. Distributional Comparison

To test whether cleaning materially altered county-level metric distributions, the study evaluated the null hypothesis:

$$H_0: F_{FIX}(m) = F_{GOOD}(m) \tag{22}$$

where $F(\cdot)$ denotes the empirical cumulative distribution function of metric m. The distributional comparison applied two complementary statistical tests:

(1) Welch's t-test, which is parametric, where

 doi:10.20944/preprints202603.0832.v1

$$t = \frac{\bar{m}_{FIX} - \bar{m}_{GOOD}}{\sqrt{\frac{s^2_{FIX}}{n_{FIX}} + \frac{s^2_{GOOD}}{n_{GOOD}}}} \quad (23)$$

with $\bar{m}$ representing the sample means, s^2 the sample variances, and n the sample sizes. This test does not assume equal variances and evaluates equality of means under approximate normality.

(2) Mann-Whitney U (MWU) test, which is nonparametric, where

$$U = \min(U_1, U_2) \quad (24)$$

with U_1 and U_2 defined as rank-based statistics computed from pooled observations. This test evaluated whether one distribution tended to produce larger values than the other without assuming normality. Applying both tests ensured robustness to heavy-tailed behavior; a statistically significant result ($p < 0.05$) indicated rejection of H_0, confirming that the cleaning strategy altered the metric distribution.

3.6.3. Relative and Ranking Effects

Beyond distributional shifts, the study evaluated proportional changes in median values where

$$R_{med} = \frac{\text{Median}_{FIX}}{\text{Median}_{GOOD}} \quad (25)$$

and state-level rank changes where

$$\Delta Rank_s = Rank_s^{FIX} - Rank_s^{GOOD} \quad (26)$$

where ranks computed after aggregating county metrics to the state level. These comparative statistics measured how the multistage cleaning process altered prioritization signals with direct implications for funding allocation. The impact assessment distinguished three possible outcomes:

1. No distributional difference but ranking shifts that indicate reallocation effects without mean distortion.
2. Significant distributional shifts, which would indicate systematic bias in baseline cleaning.
3. Denominator-driven divergence, which would indicate exposure mismeasurement under the GOOD approach.

By formalizing the comparison across numerator corrections (incident reassignment) and denominator corrections (crossing reconciliation and spatial infrastructure density), this framework quantified the methodological contribution of structured data cleaning relative to conventional record dropping (GOOD).

4. Results

The following subsections present data reconciliation outcomes, incident rate distributions, and the statistical impact of the hierarchical data cleaning pipeline.

4.1. Data Cleaning Outcomes

This subsection quantifies the effect of the multistage reconciliation framework described in Section 3. The objective was to document how many records were resolved at each stage and how the residual error structure evolved over time.

4.1.1. Distribution of Fix Types

Table 1 summarizes the number of incident records resolved and the cumulative percentage (C) of the initial BAD subset (9,425 records) resolved at each stage S of the cleaning hierarchy. Figure 2 displays the relative distribution. Stage 1 placed 250,090 records into D_1^{good}; fewer than 4% of records

 doi:10.20944/preprints202603.0832.v1

required correction. County spelling correction via RapidFuzz (CNTY_SP) accounted for 74% of reconciled records in the initial BAD subset.

Table 1. Results of the Multistage Data Cleaning Pipeline.

S	GOOD	BAD	C	Fix Type	Description
1	250,090	9,425		OK	409 non-CONUS states removed. Less than 4% bad.
2	29	9,396	0.3%	DC	Deterministic correction: 29 DC.
3	6,977	2,419	74.0%	CNTY_SP	Fuzzy county name correction for matching FIP5.
4	2,168	251	23.0%	XID	Crossing ID matching.
5	127	124	1.3%	STATION	Nearest station plus state matching.
6	37	85	0.4%	MULTI_SG	Weighted multi-signal matching. Dropped 2 in Canada.
7	65	20	0.7%	MSig_Search	Manual search partial strings with Corpus viewer.
8	20	0	0.2%	GAI	GAI with internet search verification.
9	249,676	3	100.0%		Aggregated good data and dropped 3 Alaska/Hawaii.

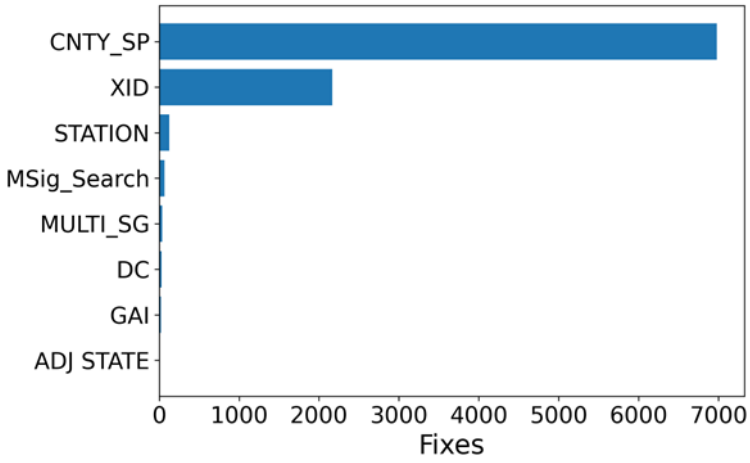

Figure 2. Distribution of the type of fix.

This result confirmed that orthographic variation was a common data-entry artifact even when FIP5 codes were valid, motivating explicit fuzzy correction as a standard pipeline stage. Infrastructure-linked corrections through XID and nearest-station matching resolved additional records with high confidence. Weighted multi-signal inference addressed a smaller but nontrivial fraction of records characterized by compound inconsistencies. Manual multi-signal search and controlled GAI inference were required only for a very small residual subset. This result was consistent with the intended hierarchical pipeline design.

4.1.2. Temporal Distribution of Data Quality

Figure 3 shows the distribution of reconciliation types across incident years. Earlier years showed higher concentrations of deterministic and spelling corrections. In particular, the five years prior to 1980 accounted for 27.15% of spelling corrections, whereas the last five years of the series accounted for only 4.6%. This pattern reflected legacy reporting systems and manual data-entry practices that predominated before electronic submissions. The relative frequency of complex multi-signal and manual corrections declined in later years. This pattern was consistent with improvements in electronic reporting and standardized identifier use.

 doi:10.20944/preprints202603.0832.v1

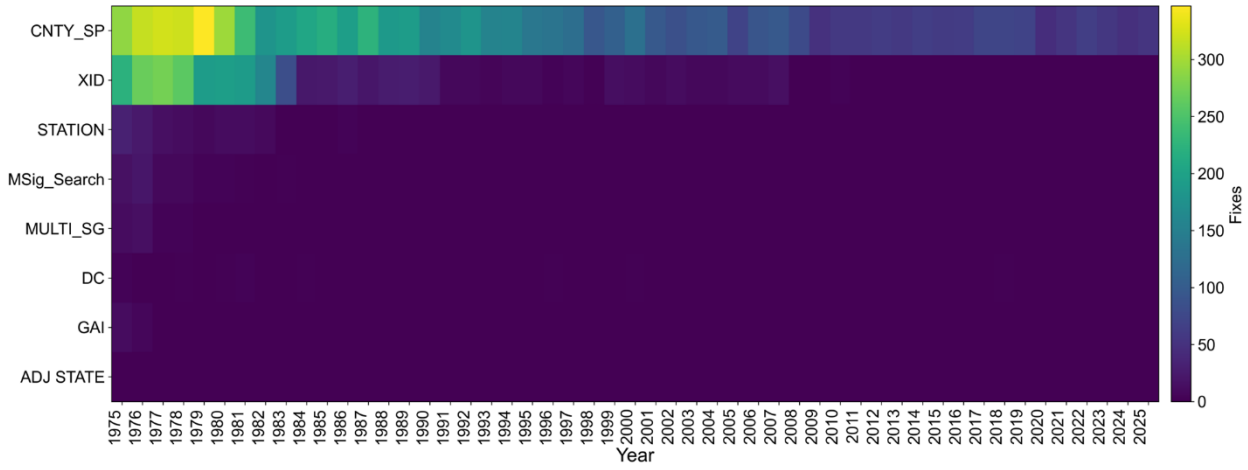

Figure 3. Heatmap of fix-type frequency by incident year; color intensity encodes count of corrections.

The temporal pattern indicates that data quality issues were not uniformly distributed across the study period. Consequently, a simple strategy of dropping inconsistent records would disproportionately affect earlier years and distort long-term trend analysis. No records remained unresolved after all reconciliation stages. The hierarchical reconciliation framework achieved complete geographic recovery while maintaining controlled acceptance thresholds.

4.1.3. GAI-Assisted Edge Cases

The final stage of the reconciliation framework applied GAI-assisted inference to a very small subset of 20 records that resisted deterministic alignment, fuzzy spelling correction, infrastructure linkage, station matching, and weighted multi-signal scoring. Although small in number, these cases were structurally complex and analytically consequential. Table 2 presents the cases in which all three platforms converged, along with the best supporting reasoning. First, several records contained internally inconsistent state–county combinations, where the county and city fields clearly aligned with a different state than the one reported. For example, “Waukesha” and “New Berlin” uniquely identify Waukesha County, Wisconsin, indicating that the listed state “Nebraska” was a mis-keyed entry. These cases could not be resolved by deterministic or fuzzy methods because the conflicts were cross-field rather than orthographic. Second, some records referenced well-known localities or rail-served communities that function as neighborhoods, industrial areas, or unincorporated places rather than incorporated municipalities. Examples include Curtis Bay (Baltimore city), Kernstown (Virginia), and Cumbo (West Virginia). These localities do not always appear in standardized city–county lookup tables. This constraint reduced the effectiveness of automated candidate restrictions. Third, phonetic distortions combined with county–city misassignment created compound ambiguity. Examples such as “PAMSEY” for Ramsey County or “SCAGETT” for Skagit County required recognition of phonetic similarity alongside contextual station or locality cues. Whereas fuzzy matching could detect single-field similarity, it could not reliably resolve which geographic hierarchy level was mis-keyed. Fourth, some cases involved domain-specific industrial references (e.g., bentonite spur, Baroid plant) associated with mineral extraction regions. These require knowledge of rail-served industrial geography that is not directly encoded in administrative boundary tables. In each case, the researcher treated GAI outputs strictly as hypotheses and accepted candidates only after independent verification against authoritative geographic references; all 20 proposed corrections were confirmed.

Although limited in number, these edge cases illustrate an important property of large administrative datasets: rare records often contained the most structurally complex errors. Resolving them ensured that geographic reconciliation was complete and that no systematic bias remained concentrated in specific industrial or rural regions.

 doi:10.20944/preprints202603.0832.v1

Table 2. The Best GAI Hypothesis for the 20 Edge Cases.

Quadplet	Best Reasoning from Multiple GAI Platforms
nebraska, waukesha, w berlin, new berlin	"Waukesha" and the station/city "New Berlin" both align with New Berlin, Wisconsin in Waukesha County, so "Nebraska" is a mis-keyed state.
ohio, missing, geauga lake oh, geauga lake oh	"Geauga Lake" is a known place name and former amusement park location in Geauga County, Ohio, so the county is missing rather than misidentified.
minnesota, pamsey, missing, daytons bluff	"PAMSEY" is a phonetic misspelling of Ramsey, and Dayton's Bluff is a well-known St. Paul neighborhood within Ramsey County, Minnesota.
minnesota, lesueur, fairbault, fairbault	"FAIRBAULT" is a misspelling of Faribault, a city and rail location in Rice County, Minnesota, so the listed Le Sueur county is incorrect.
maryland, missing, missing, penn mary md	Penn Mar is a border community on the Maryland–Pennsylvania line, located in Washington County, Maryland, commonly referenced in rail and station records as "Penn Mary."
maryland, missing, missing, stonehouse	Stonehouse is an unincorporated rail-adjacent locality in Harford County near Aberdeen Proving Ground, indicating missing county and city fields rather than a state error.
maryland, missing, missing, curtis bay	Curtis Bay is a waterfront industrial neighborhood within Baltimore city, Maryland, which functions as a county-equivalent with its own FIPS code.
minnesota, gallatin, logan, logan	"Gallatin" is a Minnesota city misused as a county, while Logan is a city and township located in Redwood County, Minnesota, indicating county misassignment.
virginia, missing, missing, kernstown	Kernstown is a historic community and rail location near Winchester, which lies in Frederick County, Virginia, indicating missing county and city fields.
west virginia, missing, missing, cumbo	Cumbo is an unincorporated coal community in McDowell County, West Virginia, commonly referenced by rail operations despite missing county and city fields.
wyoming, campell, missing, donkey creek	"CAMPELL" is a misspelling of Campbell, and Donkey Creek is a known oil and rail-served area in Campbell County, Wyoming, confirming the county and state.
wisconsin, waupoca, manowa, manowa	"WAUPOCA" is a phonetic misspelling of Waupaca, and Manawa is a city located within Waupaca County, Wisconsin, confirming the correct county and state.
west virginia, mc dowell, anawait, anawait	"ANAWAIT" is a phonetic misspelling of Anawalt, a coal-town community in McDowell County, West Virginia, confirming the county and state.
washington, scagett, missing, fir	"SCAGETT" is a phonetic misspelling of Skagit, and "FIR" indicates Fir Island, which lies entirely within Skagit County, Washington.
washington, louis, chehleis, chehleis	"CHEHLEIS" is a phonetic misspelling of Chehalis, which is the county seat of Lewis County, Washington, confirming the county and state.
south dakota, crook, baroid plant at bent, bentonite spur	Crook County is in Wyoming, not South Dakota, and bentonite plants and spurs are characteristic of Crook County, Wyoming, rail-served mineral extraction sites.
maryland daviess pattonsburg pottonsburg	The city of Pattonsburg (misspelled as POTTONSBURG in one entry) is located in Daviess County, Missouri, not Maryland.
louisiana winneshiek calmer calmar	The city of Calmar (misspelled as CALMER in one entry) is located in Winneshiek County, Iowa, not Louisiana.
louisiana terrebone colley colley	Colley is a community located within Terrebonne Parish, Louisiana. The entry "terrebone" is a misspelling of the parish name.
kentucky ibell page page	The community of Page is located in Bell County, Kentucky. The entry "IBELL" is a misspelling of the county name Bell.

4.2. Incident Rate National Distributions

This subsection describes the empirical distributions of the four county-level metrics constructed after reconciliation and crossing count accounting: AIPM, AIPC, CPHSM, and CPM. Figure 4 presents histograms, fitted theoretical distributions, and goodness-of-fit statistics for all four county-level metrics. The figure insets report summary statistics for the FIX dataset. Compared to the K-S statistic, the A-D statistic assigns greater emphasis to discrepancies in the upper and lower tails of the distribution. This complement is particularly relevant for county-level rate metrics, where extreme

 doi:10.20944/preprints202603.0832.v1

counties represent substantively important safety outliers. The distributions of all four metrics exhibited pronounced right-skewness. Most counties cluster at relatively low values, with a long upper tail representing a smaller subset of high-rate counties.

AIPM: Figure 4a shows the AIPM histogram, fitted with an exponential distribution (loc = 0.8765, scale = 1455). The K-S statistic is D = 0.0488 with p = 1.44 ×10^{-5} (n = 2479). The exposure threshold excluded counties with population < 0.01 million (kept = 2670; dropped = 438). The fitted exponential curve tracks the descending right tail in the histogram. The A-D statistic equaled A^2 = 11.5506 (p ≈ 1×10^{-27}), indicating statistically detectable deviation under large-sample conditions. The large A-D statistic reflected minor tail departures rather than central misfit. This is consistent with the exponential curve closely tracking the dominant right-tail decay in the histogram.

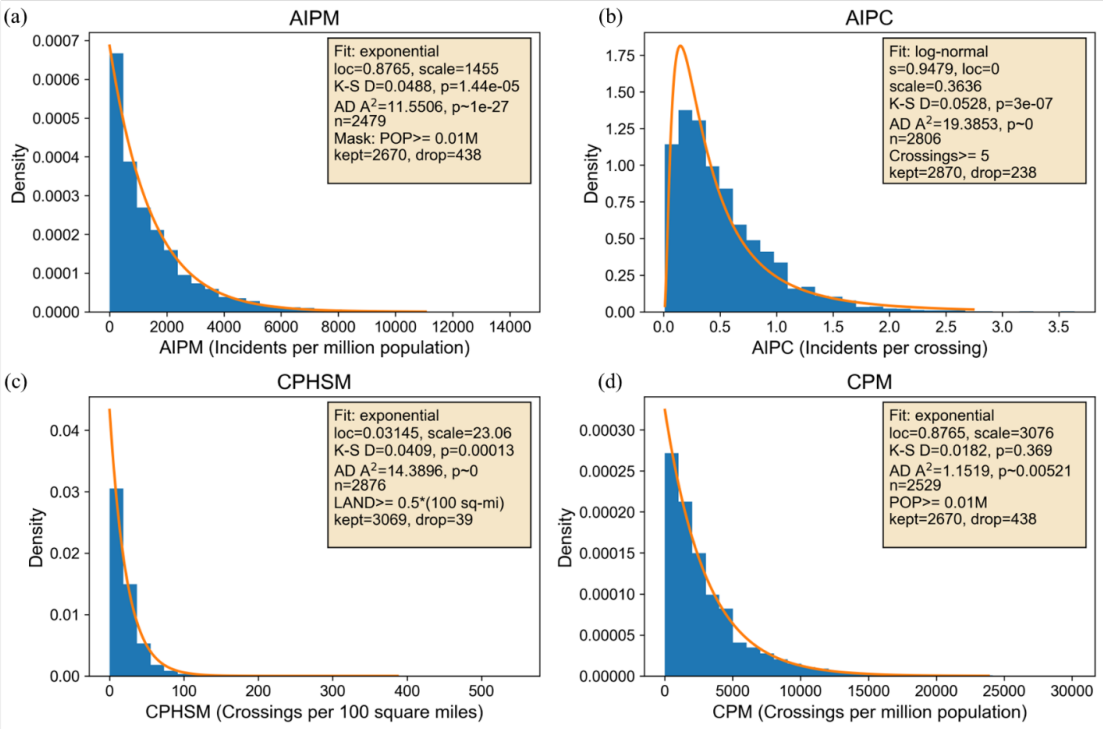

Figure 4. Statistical distributions of the fixed records.

AIPC: Figure 4b shows the AIPC histogram. It is best approximated by a log-normal distribution with shape parameter s = 0.9479 and scale = 0.3636. The K-S statistic is D = 0.0528 with p = 3.00×10^{-7} (n = 2806). The masking threshold excluded counties with fewer than five crossings (kept = 2870; dropped = 238). The histogram showed a concentration near lower values with gradual tapering across higher incident-per-crossing values. The A-D statistic equals A^2 = 19.3853 (p ≈ 0), confirming detectable deviations in the distribution tails. Given the pronounced skewness and heavier upper tail of incident-per-crossing rates, the A-D result indicates that the log-normal form captures the central mass well but does not perfectly represent extreme counties.

CPHSM: Figure 4c shows the CPHSM histogram fitted with an exponential distribution (loc = 0.03145, scale = 23.06). The K-S statistic equals D = 0.0409 with p = 0.0001298 (n = 2,876). The masking threshold excluded counties with land area below 0.5×100 square miles (kept = 3,069; dropped = 39). The distribution displayed a steep decline from low crossing-density counties toward a sparse upper tail. The A-D statistic equals A^2 = 14.3896 (p ≈ 0), again reflecting sensitivity to upper-tail density. The exponential specification captures the rapid decay from low-density counties, while tail-weighted deviations are primarily driven by a small subset of high-density outliers.

CPM: Figure 4d shows the CPM histogram, which best follows an exponential distribution (loc = 0.8765, scale = 3,076). The K-S statistic is D = 0.0182 with p = 0.3694 (n = 2,529). The same population mask as AIPM applied (kept = 2,670; dropped = 438). CPM exhibits the smallest D value (0.0182) and

 doi:10.20944/preprints202603.0832.v1

fails to reject the exponential fit at conventional significance levels. The A-D statistic equals $A^2 = 1.1519$ ($p \approx 0.00521$), indicating comparatively weak deviation relative to the other metrics. Consistent with the small K-S value (D = 0.0182), CPM exhibits the closest adherence to the exponential functional form across both central and tail regions.

The A-D statistics confirmed that tail deviations were statistically detectable under large sample sizes, yet the fitted distributions captured the dominant structural decay patterns evident in the histograms. The combined K-S and A-D diagnostics indicated that deviations arose primarily from extreme counties rather than systematic central misfit. These empirical distributional forms therefore provided a defensible parametric basis for subsequent FIX–GOOD distributional comparisons.

4.3. Incident Rate Spatial Distributions

This subsection presents the geographic distribution of AIPM after completion of the multistage reconciliation and accounting procedure for the number of crossings.

4.3.1. Counties

Figure 5 displays the county-level AIPM values using Jenks natural breaks with four classes. Higher AIPM classes concentrated in the central United States. Concentrations appeared in Kansas, Oklahoma, Arkansas, Missouri, Mississippi, Louisiana, Iowa, Nebraska, and parts of Texas. Additional counties with elevated values are visible in portions of West Virginia and surrounding Appalachian areas. In contrast, many counties in the Northeast, upper Midwest metropolitan areas, and large portions of the Mountain West and Pacific regions fell within the lowest class or reported no incidents. Western states showed large contiguous areas of low AIPM values, particularly in the Mountain West and Southwest. Isolated moderate values appear along selected corridor-aligned counties in Arizona, New Mexico, Colorado, and Texas.

Table 3 lists the 10 counties with the highest AIPM values and the mean of their 2010–2020 census estimates. These top-ranked counties lie in the Central Plains, lower Mississippi region, and Appalachian coal region. Overall, the county-level map revealed distinct regional concentrations. Higher AIPM values were concentrated in central and southern interior regions, and lower values were prevalent across densely populated coastal states and much of the western United States.

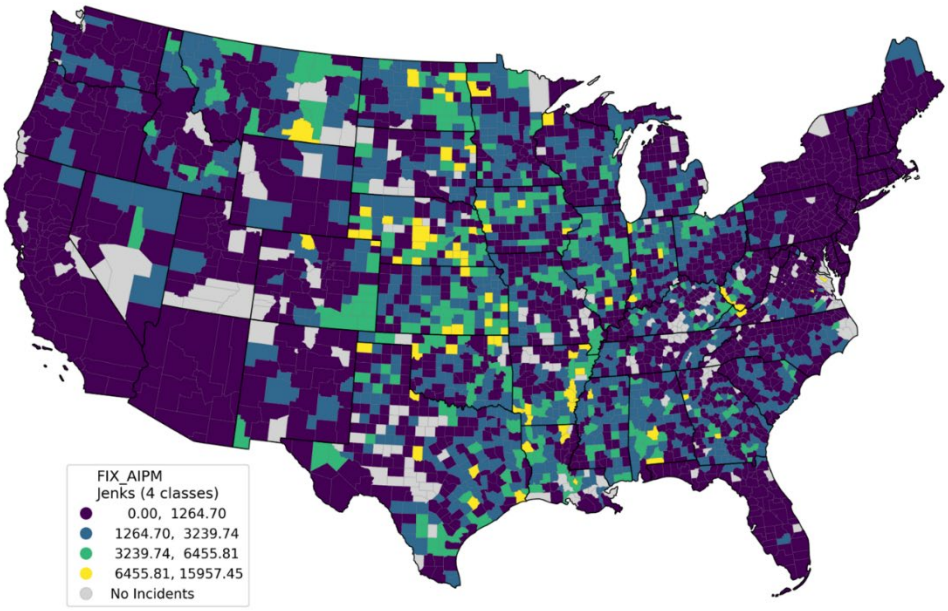

Figure 5. Spatial distribution of AIPM by county.

 doi:10.20944/preprints202603.0832.v1

Table 3. The top 10 Counties by AIPM.

State-CFIP	County	Population Mean	AIPM
KS-205	Wilson County	6,016	15,957
AR-073	Lafayette County	6,500	15,077
TX-395	Robertson County	10,541	14,325
AR-095	Monroe County	6,803	13,670
AR-081	Little River County	12,131	12,777
TX-111	Dallam County	5,245	11,440
NE-059	Fillmore County	3,963	11,355
LA-095	St. John the Baptist Parish	33,634	11,239
TX-197	Hardeman County	4,410	10,884
WV-059	Mingo County	25,166	10,848

4.3.2. States

Figure 6 and Figure 7 present AIPM aggregated at the state level. At this level, AIPM was calculated by summing total incidents and total population within each state and computing incidents per million population. Figure 6 ranks states by AIPM. The highest values occurred in Arkansas, Alabama, Louisiana, Mississippi, Indiana, Nebraska, North Dakota, Iowa, and Kansas. These states formed a contiguous corridor from the lower Mississippi Valley through the Central Plains and Midwest. Intermediate-ranked states include Illinois, Ohio, Kentucky, Wisconsin, South Dakota, Oklahoma, Minnesota, and Montana. Lower AIPM values appear in the Northeast (e.g., Massachusetts, New York, Rhode Island, Connecticut), selected Mid-Atlantic states, and portions of the western United States including California and Nevada. The District of Columbia and Rhode Island occupy the lowest positions in the ranking.

 doi:10.20944/preprints202603.0832.v1

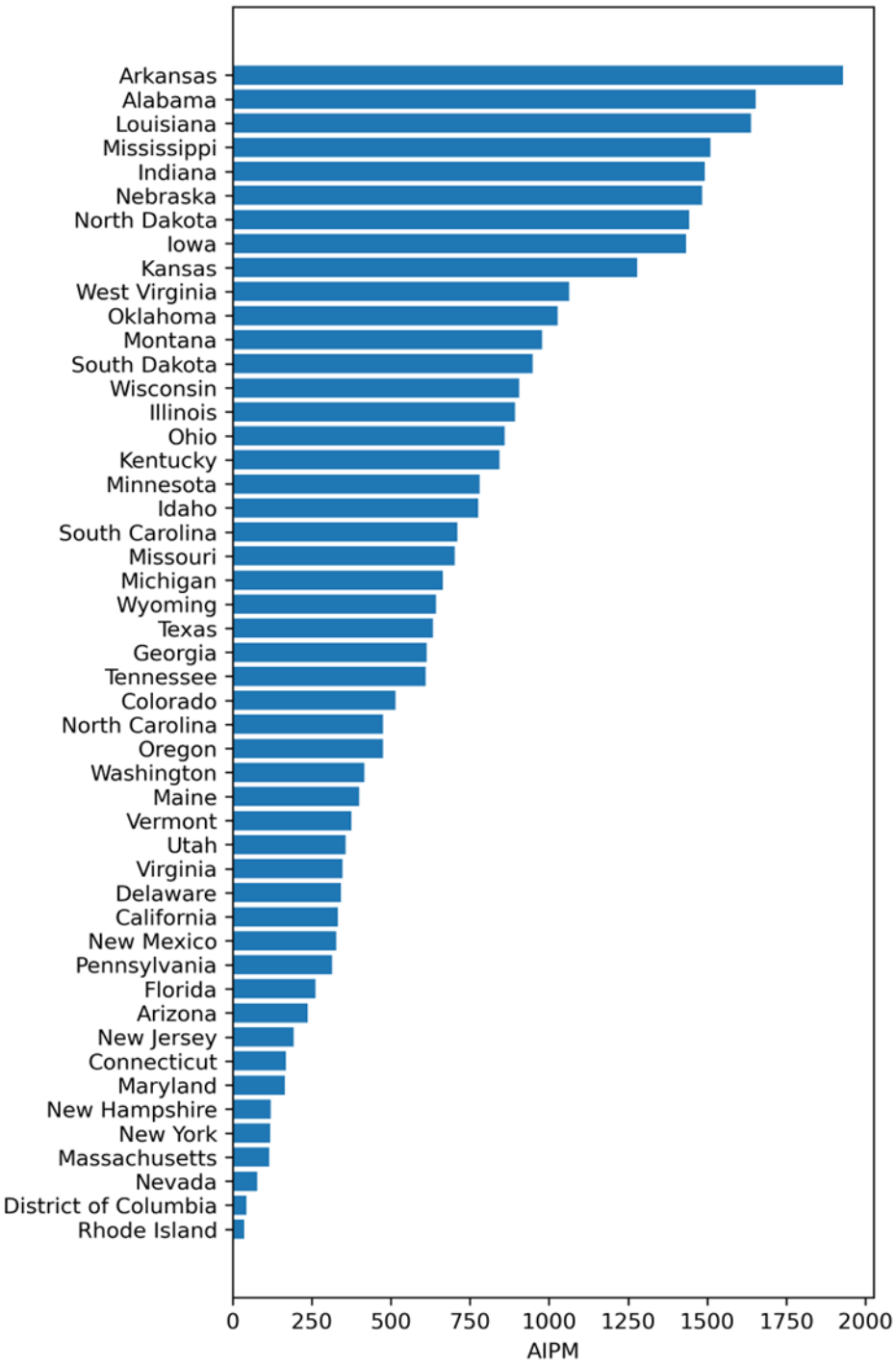

Figure 6. State rank by AIPM.

Figure 7 displays the same state-level AIPM values using Jenks natural break classification. The Jenks boundaries were data-driven; they are not policy thresholds. The Jenks classification method minimizes within-class variance while maximizing between-class variance. This optimization is well-suited for the heavy-tailed distributions where equal-interval or quantile schemes would either obscure the concentration of high-rate counties or misrepresent the sparsely populated upper tail. The map shows that the highest Jenks classification category was concentrated in the lower Mississippi Valley and Central Plains. Moderate classes extend across portions of the Midwest and South. The lowest class is concentrated in the Northeast and parts of the western United States. The

 doi:10.20944/preprints202603.0832.v1

state-level aggregation preserves the broad regional structure observed at the county scale while reducing local variability.

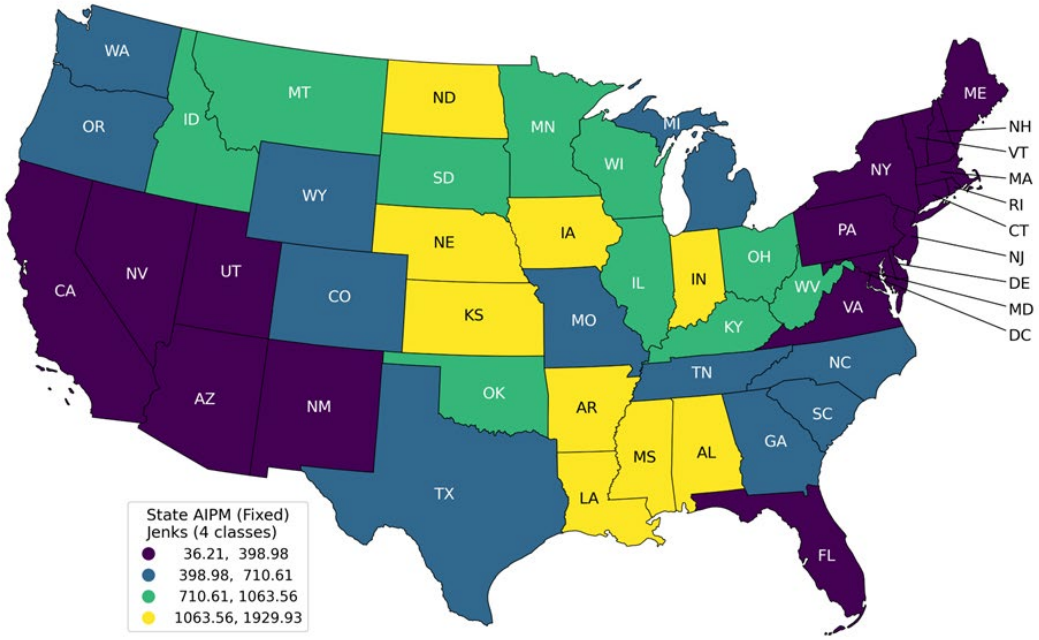

Figure 7. AIPM by state with Jenks Natural Break Classification.

4.4. National Temporal Trend

Figure 8 presents the Bai–Perron segmented annual incident trend with regime-specific functional forms. Table 4 summarizes the evidence for selecting two trend breakpoints. These were in 1980 and 2001, which yielded three distinct phases.

Table 4. Model Selection Evidence Based on BIC Changes.

Comparison	ΔBIC	Interpretation
0 → 1 break	−74.7	Very strong evidence for structural change
1 → 2 breaks	−45.6	Very strong evidence for a third regime
0 → 2 breaks	−120.3	Overwhelming support for three phases

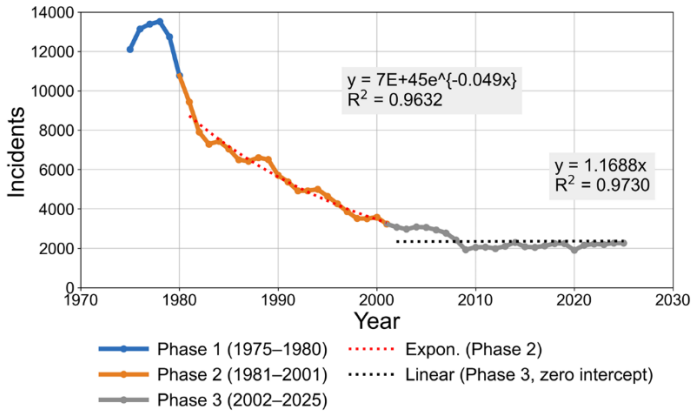

Figure 8. Long-term temporal trend of annual incidents.

4.4.1. Phase 1 (1975–1980): Transitional Volatility

 doi:10.20944/preprints202603.0832.v1

The early period exhibited elevated incident levels exceeding 13,000 annually, followed by an abrupt decline by 1980. This short interval does not demonstrate a stable monotonic structure and is therefore presented descriptively without parametric fitting. The variability reflected transitional conditions rather than a sustained safety regime.

4.4.2. Phase 2 (1981–2001): Systemic Exponential Decline

An exponential model characterized the dominant decline phase:

$$y = 7 \times 10^{45}\, e^{-0.049x}, (R^2 = 0.9632). \quad (27)$$

The high R^2 = 0.9632 confirmed that the decline followed a proportional decay process. The exponent -0.049 implied an average annual reduction rate of approximately 4.9%. The exponential specification was appropriate because reductions were multiplicative rather than linear, with each year's reduction scaling relative to the current incident level. The strong fit ($R^2 \approx 0.96$) confirmed a highly stable structural decline over two decades.

4.4.3. Phase 3 (2002–2025): Stabilization Plateau

The post-2001 regime exhibited markedly slower change. A linear model provided an adequate representation as

$$y = 1.1688x, (R^2 = 0.9730). \quad (28)$$

The near-horizontal trend indicated that annual incident levels fluctuated within a narrow band, averaging 2,337 incidents per year with a marginal annual increase of approximately 0.05%. The minimal slope relative to the magnitude of counts confirmed structural stabilization. Whereas Phase 2 achieved rapid proportional reductions, Phase 3 reflected diminishing marginal safety gains. The structural transition in 2001 therefore marked the shift from accelerated systemic improvement to long-run equilibrium behavior.

4.5. State-Wise Temporal Trend

Figure 9 presents the annual evolution of AIPM by state from 1975 through 2025. The heatmap ordered states vertically by overall magnitude. The heatmap positioned higher long-run AIPM values toward the top. During the late 1970s and early 1980s, the highest color intensities concentrated in Arkansas, Alabama, Louisiana, Mississippi, Indiana, Nebraska, North Dakota, Iowa, and Kansas. These states display consistently elevated AIPM values during the early years of the series.

 doi:10.20944/preprints202603.0832.v1

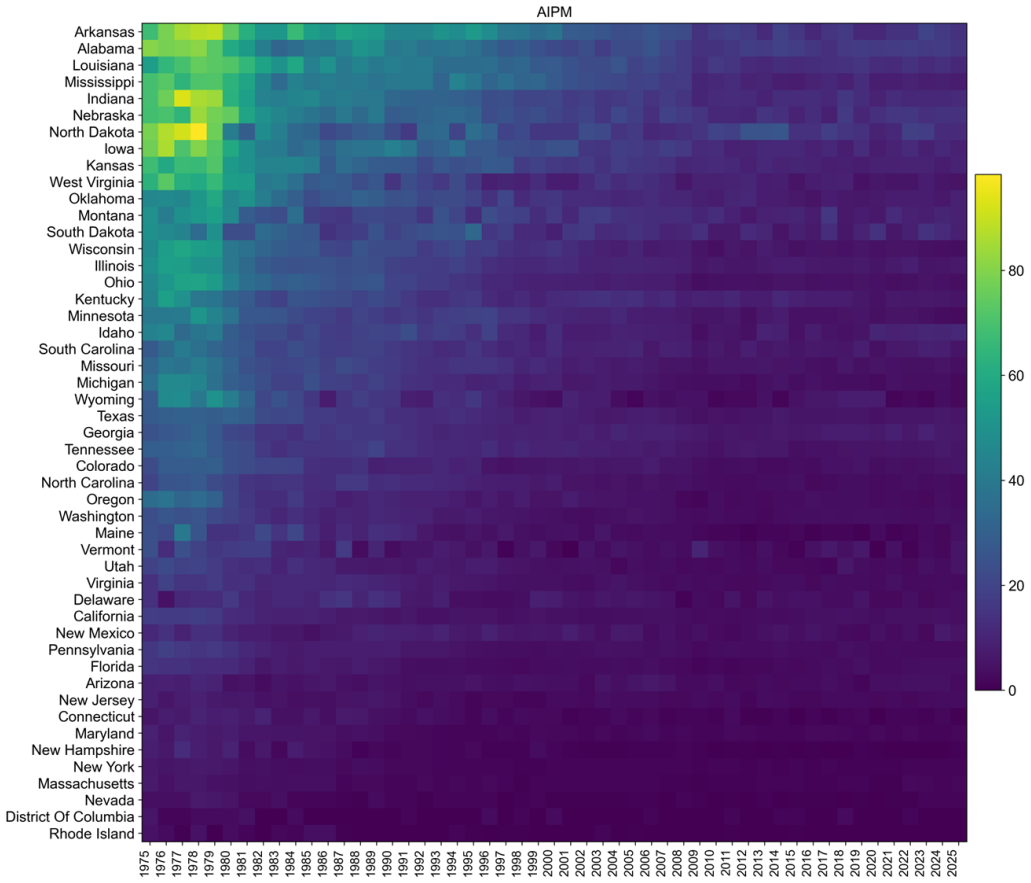

Figure 9. Long-term AIPM by state.

From approximately 1980 through the 1990s, the heatmap shows a broad and sustained reduction in intensity across nearly all states. The transition from higher to lower color intensity occurs progressively over time. By the early 2000s, most states exhibit substantially reduced AIPM levels relative to the late 1970s. After approximately 2001, the heatmap stabilized, with most states displaying relatively uniform low-intensity coloring and limited year-to-year variation; the contrast between states was smaller than in earlier decades.

Despite overall convergence toward lower AIPM values, relative differences persisted. States in the Central Plains and lower Mississippi Valley continued to occupy higher positions in the ordering. Northeastern states, including Massachusetts, New York, New Jersey, Rhode Island, and Connecticut, consistently appeared near the bottom of the ordering throughout the series. Western states in the Mountain West and Pacific regions showed moderate early values followed by steady reduction and stabilization.

4.6. Data Cleaning Impact

4.6.1. Statistical Differences

Figure 10 contrasts the four county-level metrics under the GOOD and FIX strategies.

 doi:10.20944/preprints202603.0832.v1

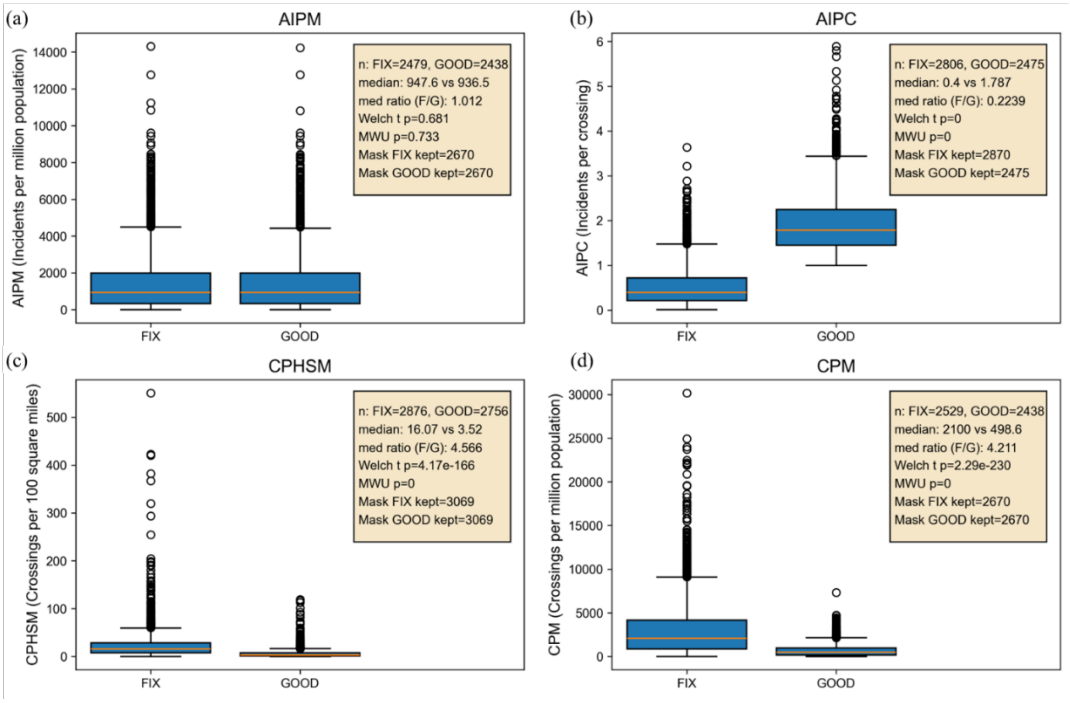

Figure 10. Distribution of the four normalized HRGC metrics.

The GOOD dataset retained only internally consistent incident records and derived crossing counts from those retained records. The FIX dataset incorporated multistage FIP reconciliation and crossing union construction. The boxplot insets report medians, sample sizes, masking thresholds, and Welch's t-test and MWU test results.

Figure 10a shows that the median AIPM was 947.6 for FIX and 936.5 for GOOD (median ratio = 1.012); Welch's test yielded p = 0.681 and the MWU test yielded p = 0.733, neither rejecting distributional equality at the 0.05 level. The masking threshold removed counties with populations below 0.01 million (kept = 2,670).

Figure 10b shows that the median AIPC was 0.400 for FIX and 1.787 for GOOD (median ratio = 0.22). Welch's test yielded $p \approx 0$, and the MWU test yielded $p \approx 0$. Both tests therefore rejected equality of distributions. The procedure masked counties with fewer than five crossings (kept = 2,870) for consistency with the stability of their distributional fitting.

Figure 10c shows that the median CPHSM was 16.07 for FIX and 3.52 for GOOD (median ratio = 4.566). Welch's test yielded $p = 4.17 \times 10^{-166}$, and the MWU test yielded $p \approx 0$. The procedure masked counties with land area below 0.5 × 100 square miles (kept = 3,069) for consistency with the stability of their distributional fitting.

Figure 10d shows that the median CPM was 2,100 for FIX and 498.6 for GOOD (median ratio = 4.211). Welch's test yielded $p = 2.29 \times 10^{-230}$, and the MWU test yielded $p \approx 0$. The population mask (≥ 0.01 million) kept 2,670 records.

AIPM exhibited distributional stability under both parametric and nonparametric tests because both datasets shared the same external population denominator. In contrast, AIPC, CPHSM, and CPM showed statistically significant differences under both parametric and nonparametric tests ($p \approx 0$) because they depended on the reconstructed crossing universe. The magnitude of median differences was largest for the crossing-based metrics (CPHSM and CPM), followed by AIPC.

4.6.2. State Rank Change

Figure 11 displays changes in state-level AIPM ranking when moving from the GOOD dataset to the FIX dataset.

 doi:10.20944/preprints202603.0832.v1

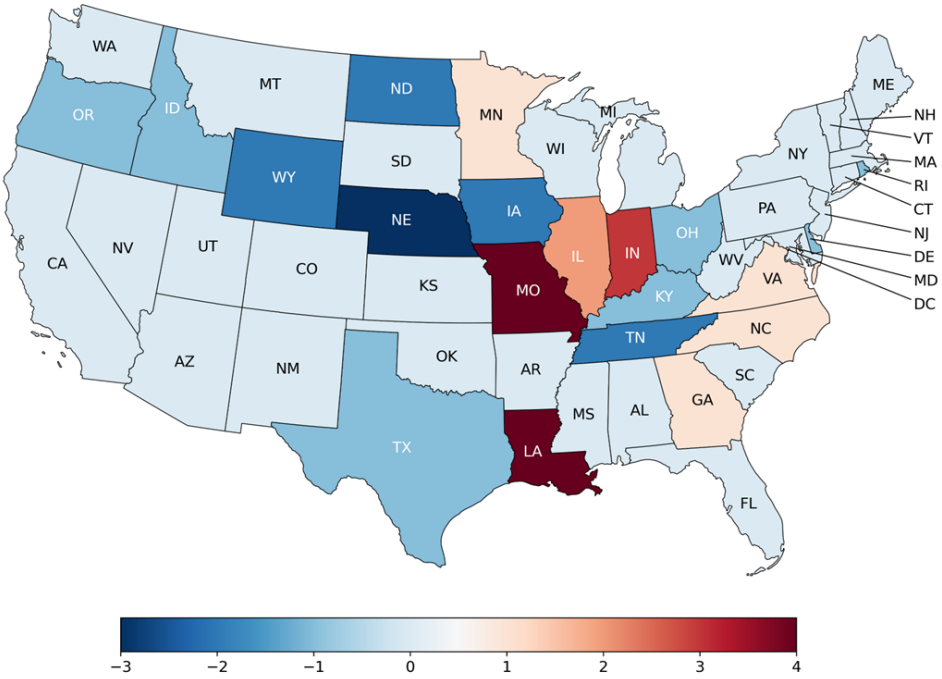

Figure 11. State AIPM rank change going from GOOD to FIX scenarios.

Positive values indicate that a state moved upward in relative AIPM ranking under the FIX scenario. Negative values indicated a downward movement. The largest upward rank shifts are observed in Louisiana and Missouri, followed by Indiana and Illinois. Additional upward movements appear in selected states within the lower Mississippi Valley and Midwest. The largest downward rank shifts occur in Nebraska, Wyoming, North Dakota, Tennessee, and Iowa. These states move lower in relative AIPM position under the FIX scenario compared with the GOOD scenario. County name corrections that recovered incidents previously misattributed to incorrect states drove the rank changes. This outcome confirmed that numerator redistribution—not denominator expansion—produced the AIPM shifts. The analysis showed minimal rank changes in several northeastern states, where ordering remains largely stable between scenarios. Coastal states and selected western states also exhibit limited movement.

Rank changes concentrated regionally. Positive shifts are concentrated in portions of the Mississippi River Basin and Midwest. Negative shifts are more visible across parts of the Great Plains and Mountain West. Although the two-sample tests detected no statistically significant distributional difference for AIPM, Figure 11 shows that relative state ordering changed under the FIX scenario. The magnitude of these shifts varied by state, with some states moving several positions in either direction.

5. Discussions

This study examined how multistage geographic reconciliation and denominator reconstruction changed the interpretation of long-run HRGC safety patterns. This section integrates cleaning outcomes, distributional structure, spatial patterns, temporal dynamics, and statistical impact into a unified interpretation.

5.1. Data Integrity and Longitudinal Validity

Table 1 and Figure 2 showed that the data cleaning pipeline resolved most corrections deterministically or through structured fuzzy matching. Only a small residual required manual or GAI-assisted inference. Large administrative safety datasets can therefore be rehabilitated through auditable, reproducible logic rather than subjective editing. Figure 3 demonstrated that geographic

 doi:10.20944/preprints202603.0832.v1

defects concentrated in earlier decades; a record-dropping strategy would therefore have removed a disproportionate share of historical observations, biasing long-run trend analysis. The edge cases summarized in Table 2 revealed four distinct error categories: cross-field state–county conflicts, phonetic distortions (e.g., "PAMSEY" for Ramsey, "SCAGETT" for Skagit), references to unincorporated or industrial localities absent from administrative lookup tables, and domain-specific rail-industry nomenclature. These patterns constituted systematic structural artifacts—not random noise—concentrated in earlier years when manual data entry predominated.

5.2. Distributional Structure and Modeling Design

Figure 4 showed strong right-skewness across the four metrics. AIPM, CPHSM, and CPM followed exponential-like decay, whereas AIPC followed a log-normal form. The K-S statistics confirmed that these families captured the dominant distributional shape, with CPM failing to reject the exponential fit at conventional significance levels. Heavy right tails indicate that a small subset of high-rate counties drives most of the variance. The log-normal behavior of AIPC was consistent with multiplicative crossing-level processes. Masking thresholds demonstrated that denominator instability materially affected parametric estimation. Together, these results indicate that downstream modeling should employ robust, distribution-aware methods.

5.3. Spatial Coherence and Geographic Plausibility

The county AIPM map in Figure 5 showed coherent regional clustering rather than fragmented anomalies. Elevated classes concentrate across the Central Plains and lower Mississippi corridor. Characteristics of these regions include dense Class I mainlines, heavy agricultural bulk freight, and rural HRGCs. The top 10 counties listed in Table 3 align geographically with these concentrations. State-level aggregation in Figure 6 and Figure 7 preserve this structure. States with high AIPM values formed a contiguous interior corridor. On the other hand, states with lower AIPM values concentrated in the Northeast and selected western regions. The consistency between state- and county-level results indicated that county-to-state aggregation was internally coherent after reconciliation; systematic FIP distortion would have generated implausible discontinuities across adjacent jurisdictions.

5.4. National Decline and Structural Stabilization

The fitted models of Figure 8 reveal three qualitatively different safety dynamics:
Pre-1981 volatility without a stable decline structure
1981–2001 exponential decay, representing the primary safety transformation era
Post-2001 stabilization, characterized by low variance and incremental improvement

The strong exponential fit in Phase 2 and the near-flat linear behavior in Phase 3 confirmed that the decline in HRGC incidents was not uniform across the 51-year period. Instead, the safety system progressed through a rapid reform phase followed by long-run stabilization at a lower incident baseline. These results confirm the three-phase structural narrative that Bai–Perron break detection and BIC model selection support in Table 4.

The timing of the decline was consistent with expanded federal–state crossing investment, active warning device installation, crossing consolidation, and rail system restructuring documented in prior literature [9] [44]; this study did not test these causal mechanisms directly. Specifically, the Section 130 program established by the Highway Safety Act of 1973 provided dedicated funding for crossing safety improvements, the widespread installation of active warning devices (gates and lights), and crossing consolidation efforts to remove redundant or high-risk crossings [44].

The stabilization phase suggested that remaining incidents were embedded in structural exposure constraints rather than uniformly reducible through broad interventions. The smooth exponential decay and stable plateau in Figure 8 provided an indirect internal validity check on the reconciled dataset, as systematic geographic misassignment would have introduced artificial

 doi:10.20944/preprints202603.0832.v1

discontinuities inconsistent with $R^2 \approx 0.97$. This result motivates future research to examine whether residual Phase 3 incidents differ systematically from those that declined most sharply during Phase 2—for instance in crossing type, warning device presence, and proximity to population centers. The warning device composition of incidents shifted materially across the three phases identified in Section 3, and it provided direct empirical evidence relevant to the crossing-type question posed above. Table 5 reports the percentage of incidents within each phase attributable to four warning device categories: crossing gates, crossbucks, flashing light signals (FLS), and stop signs.

Table 5. Percentage of Incidents Within the Three Phases Involving Safety Mechanisms.

Phase	Gates	Crossbuck	FLS	Stop Signs
1	9.5	46.7	26.0	2.0
2	18.4	44.3	19.5	3.3
3	42.0	23.1	9.2	15.9

Gate-protected crossings accounted for 9.5% of Phase 1 incidents but rose to 42.0% in Phase 3, while crossbuck- and FLS-associated incidents declined from 46.7% and 26.0% in Phase 1 to 23.1% and 9.2% in Phase 3, respectively. Stop-sign crossings increased from 2.0% in Phase 1 to 15.9% in Phase 3. These compositional shifts reflected the widespread installation of active warning devices documented in prior literature [44], which progressively reduced incident exposure at the least-protected crossings during Phase 2. The residual Phase 3 incident pool therefore concentrated at crossings with either the highest protection level—gates—or the least regulated access category—stop signs—suggesting that continued reductions depend on targeted strategies rather than broad device-upgrade programs.

5.5. State Convergence and Persistent Ordering

Figure 9 shows that in the late 1970s nearly all states experienced elevated AIPM followed by broad decline. By the early 2000s, convergence toward lower levels is visible across most states. After 2001, stabilization mirrors the national pattern in Figure 8. Despite convergence, relative ordering persisted. Central interior states remained near the top of the AIPM distribution, while northeastern states remained near the bottom. This persistence indicated that structural exposure differences remained after decades of improvement.

5.6. Denominator Integrity as the Primary Analytical Driver

Figure 10 isolated the methodological contribution. AIPM medians under FIX and GOOD were statistically indistinguishable (Welch $p = 0.681$; MWU $p = 0.733$). In contrast, AIPC, CPHSM, and CPM differed materially and significantly. AIPM used an external census population denominator, making it invariant to crossing reconciliation. Crossing-based metrics depended on the reconstructed crossing universe. The GOOD scenario counts crossings only from retained incident records, creating an incident-conditioned denominator that shrinks with each dropped or misattributed record. The FIX scenario reconstructed crossings from the reconciled union set. The denominator therefore expanded. The GOOD scenario constructed an incident-conditioned denominator containing only crossings observed in retained incident records—a subset of the true at-grade crossing population:

$$X_c^{GOOD} = |\{XID: incident\ observed\}| \tag{29}$$

and

$$X_c^{GOOD} \subseteq X_c^{FIX}. \tag{30}$$

Therefore,

$$AIPC^{GOOD} = \frac{I_c}{X_c^{GOOD}} \geq \frac{I_c}{X_c^{FIX}}. \tag{31}$$

 doi:10.20944/preprints202603.0832.v1

Furthermore, the shifts in median values (AIPC ratio ≈ 0.22; CPHSM ≈ 4.566; CPM ≈ 4.211) quantified denominator compression under GOOD. Thus, the principal impact of this measurement was denominator integrity rather than numerator expansion.

5.7. Rank Sensitivity Despite Distributional Stability

Figure 11 showed that state AIPM rankings shifted under the FIX scenario even though distributional tests failed to reject equality. This distinction separates statistical distributional similarity from the ranking stability relevant to resource allocation. Even modest redistribution of incident attribution across counties altered comparative positioning at the state level. The rank-change map demonstrated that reconciliation affected comparative prioritization signals even when central tendencies remained stable.

5.8. Policy Relevance and Future Direction

Spatial concentration, long-run decline, state convergence, and denominator sensitivity collectively establish that exposure-based prioritization requires reconciled crossing counts. Population-normalized metrics alone do not support infrastructure-focused decision-making. The stabilization of incident rates after 2001 indicated a structural plateau, suggesting that conventional safety interventions had reached a point of diminishing marginal returns. The persistent regional differentials indicate that future reductions may depend on targeted strategies informed by crossing density, incidents per crossing, and corridor-specific exposure.

5.9. Limitations

This study faces three limitations that bound but do not invalidate the central conclusions. First, the crossing denominator suffers from left-censoring prior to approximately 1985. Crossings retired before 1975 are absent from both the incident file and the inventory extract, rendering those crossings structurally unobservable. Because the reconciled crossing universe cannot recover infrastructure that predates the federal reporting horizon, AIPC values computed for the earliest years of the series understate the true denominator. Trend comparisons involving AIPC before 1985 should therefore be interpreted with caution, and analysts extending this framework to pre-1985 corridor-level studies should apply explicit denominator-uncertainty bounds. Second, the FIX and GOOD distributional comparisons in Section 3 were conducted over partially overlapping county sets. The FIX dataset recovered records that the GOOD strategy dropped, and those recovered records are disproportionately concentrated in earlier decades and in rural, lower-population counties. This recovery pattern is not random; it reflects the systematic geographic and temporal concentration of legacy data-entry errors documented in Figure 3. The directional implication is that the GOOD strategy understates incident exposure in precisely those counties and periods where infrastructure density and historical incident burden were highest. Third, the parametric families fitted in Section 3 describe empirical regularities but do not identify underlying generative processes. Exponential and log-normal forms characterize the observed county-level distributions, but they do not establish the behavioral, operational, or regulatory mechanisms that produced those distributional shapes. This limitation bounds causal claims throughout the paper. Taken together, these three constraints are structural rather than analytical failures. None alters the central conclusion: structured geographic reconciliation and crossing denominator reconstruction materially improved the validity of exposure-based HRGC safety metrics and stabilized comparative inference across counties and states.

6. Conclusions

Highway–rail grade crossing safety remains a persistent national concern despite decades of documented decline in incident frequency. This study addressed the foundational problem of whether long-run HRGC safety metrics were materially affected by incomplete geographic

 doi:10.20944/preprints202603.0832.v1

reconciliation and inconsistent crossing denominator construction. By analyzing 51 years of national incident data (1975–2025) and integrating them with the FRA crossing inventory, this study established a reproducible framework for identifier harmonization, crossing reconstruction, and exposure-based metric validation.

The nine-stage pipeline resolved all 9,422 geographic inconsistencies, preserved fix-type provenance labels at each stage, and achieved 100% geographic recovery after CONUS filtering. Deterministic FIP code alignment and structured fuzzy matching resolved most discrepancies. Infrastructure-level linkage through XID and station matching resolved additional cases. A small residual subset required structured manual inference and controlled AI-assisted hypothesis testing with independent verification. The resulting FIX dataset provided standardized county identifiers and a reconciled crossing universe across the full 51-year horizon.

Empirical results demonstrated three central findings. First, county-level metrics exhibited consistent heavy right-skewed distributions. AIPM, CPHSM, and CPM followed exponential-like forms. However, AIPC followed a log-normal form. These distributional properties reflected structural heterogeneity across counties and justified robust statistical treatment. Second, spatial patterns were coherent and regionally structured. Elevated AIPM values concentrated in the Central Plains and lower Mississippi corridor at both county and state levels. The national time series displayed elevated incident counts in the late 1970s, a sustained exponential decline beginning around 1980, and stabilization after approximately 2001. However, this stabilization persisted at more than 2,000 incidents annually. The state-level heatmap confirmed widespread convergence across jurisdictions while preserving persistent relative ordering. Third, and most importantly, denominator integrity drove exposure-based inference. Population-normalized AIPM distributions remained statistically indistinguishable under record-dropping (Welch $p = 0.681$; MWU $p = 0.733$). In contrast, crossing-based metrics (AIPC, CPM, CPHSM) changed materially and significantly after reconstructing crossing counts from the reconciled union of incident and inventory data. Median differences exceeding four-fold for crossing density metrics quantified the distortion introduced by incident-only denominator construction. Furthermore, state ranking shifts occurred under the reconciled dataset even when distributional tests did not reject equality. This demonstrated that comparative prioritization signals were sensitive to geographic reconciliation.

The principal contribution is formalizing and quantifying the methodological impact of multistage data cleaning and crossing denominator reconstruction over a 50-year horizon. Whereas HRGC research frequently examined incident trends or risk factors, few studies documented the technical structure of data rehabilitation or measured how denominator inconsistencies propagated into exposure metrics. By isolating the impact of reconciliation relative to a conventional baseline, this study demonstrated that rigorous geographic harmonization is not ancillary preprocessing. Rather, it is integral to valid exposure assessment.

The benefits of this contribution extend to researchers, transportation agencies, and policy analysts. The cleaned and audited dataset enables more defensible county- and state-level comparisons and is suitable for public archiving to support replication by the broader HRGC research community. The documented reconciliation hierarchy provides a replicable template for other large administrative transportation datasets characterized by evolving identifiers and cross-file identifier inconsistencies. The intermediate descriptive outputs of distributional structure, spatial clustering, temporal phases, and denominator sensitivity offer actionable insight even without predictive modeling.

This study establishes that long-run HRGC trend interpretation is robust at the population-aggregate level but highly sensitive at the infrastructure-exposure level, a distinction with direct consequences for crossing-level prioritization and funding allocation. The study clarified the conditions under which simple record-dropping is adequate and the conditions under which it materially distorts inference. It also demonstrates that 50 years of national HRGC data can be reconciled into a coherent analytic framework with transparent correction provenance. Future work will use the reconciled dataset to develop predictive and spatial econometric models of crossing-level

 doi:10.20944/preprints202603.0832.v1

and county-level risk, evaluate targeted intervention strategies, and examine corridor-specific exposure dynamics. The reconciled dataset supports hazard modeling, hierarchical Bayesian estimation, spatial econometrics, and crossing-level risk stratification. Such modeling efforts depend critically on the denominator integrity and geographic consistency this study establishes.

Funding: This research was funded by the United States Department of Transportation, grant number 69A3552348308.

Data Availability Statement: This article includes the data presented in the study.

Conflicts of Interest: The author declares no conflicts of interest.

References

[1] USDOT Form 57, "Highway-Rail Grade Crossing Incident Data (Form 57)," 5 March 2026. [Online]. Available: https://data.transportation.gov/Railroads/Highway-Rail-Grade-Crossing-Incident-Data-Form-57-/7wn6-i5b9/about_data. [Accessed 5 March 2026].

[2] M. Zayandehroodi, B. Mojaradi and M. Bagheri, "Improving reliability of safety countermeasure evaluation at highway-rail grade crossings through aleatoric uncertainty modeling with machine learning techniques," *Reliability Engineering & System Safety,* vol. 261, p. 111082, 2025.

[3] S. C. Mok and I. Savage, "Why Has Safety Improved at Rail-Highway Grade Crossings?," *Risk Analysis,* vol. 25, no. 4, pp. 867-881, 2005.

[4] E. Senkondo, D. Chimba, M. Madalo, A. Yeboah and S. Blue, "Comparative Analysis of Machine Learning and Statistical Models for Railroad–Highway Grade Crossing Safety," *Vehicles,* vol. 7, no. 4, p. 163, 2025.

[5] M. U. Farooq and A. J. Khattak, "Investigating highway–rail grade crossing inventory data quality's role in crash model estimation and crash prediction," *Applied Sciences,* vol. 13, no. 20, p. 11537, 2023.

[6] P. Martins, F. Cardoso, P. Váz, J. Silva and M. Abbasi, "Performance and Scalability of Data Cleaning and Preprocessing Tools: A Benchmark on Large Real-World Datasets," *Data,* vol. 10, no. 5, p. 68, 2025.

[7] C. Cheng, L. Messerschmidt, I. Bravo, M. Waldbauer, R. Bhavikatti, C. Schenk, V. Grujić, T. Model, R. Kubinec and J. Barceló, "A General Primer for Data Harmonization," *Scientific Data,* vol. 11, no. 1, p. 152, 2024.

[8] J. Mathew, R. F. Benekohal, M. Berndt, J. Beckett and J. McKerrow, "Multi-criteria prioritization of highway-rail grade crossings for improvements: a case study," *Urban Planning and Transport Research,* vol. 9, no. 1, pp. 479-518, 2021.

[9] O. F. Abioye, M. A. Dulebenets, J. Pasha, M. Kavoosi, R. Moses, J. Sobanjo and E. E. Özgüven, "Accident and hazard prediction models for highway–rail grade crossings: a state-of-the-practice review for the USA," *Railway Engineering Science,* vol. 28, no. 3, pp. 251-274, 2020.

[10] J. Liu, X. Wang, A. J. Khattak, J. Hu, J. Cui and J. Ma, "How big data serves for freight safety management at highway-rail grade crossings? A spatial approach fused with path analysis," *Neurocomputing,* vol. 181, pp. 38-52, 2016.

[11] S. Karimi, A. Hosseinzadeh, R. Kluger, T. Wang, R. R. Souleyrette and E. Harding, "A systematic review and meta-analysis of data linkage between motor vehicle crash and hospital-based datasets," *Accident Analysis & Prevention,* vol. 197, p. 107461, 2024.

[12] A. Hosseinzadeh, A. Karimpour, R. Kluger and R. Orthober, "Data linkage for crash outcome assessment: Linking police-reported crashes, emergency response data, and trauma registry records," *Journal of Safety Research,* vol. 81, pp. 21-35, 2022.

[13] A. Soltani, J. Harrison, C. Ryder, J. Flavel and A. Watson, "Police and hospital data linkage for traffic injury surveillance: A systematic review," *Accident Analysis & Prevention,* vol. 197, p. 107426, 2024.

 doi:10.20944/preprints202603.0832.v1

[14] P. Lu and D. Tolliver, "Accident prediction model for public highway-rail grade crossings," *Accident Analysis & Prevention,* vol. 90, pp. 73-81, 2016.

[15] X. Zhou, P. Lu, Z. Zheng, D. Tolliver and A. Keramati, "Accident Prediction Accuracy Assessment for Highway-Rail Grade Crossings Using Random Forest Algorithm Compared with Decision Tree," *Reliability Engineering & System Safety,* vol. 200, p. 106931, 2020.

[16] A. Keramati, P. Lu, Y. Ren, D. Tolliver and C. Ai, "Investigating the effectiveness of safety countermeasures at highway-rail at-grade crossings using a competing risk model," *Journal of Safety Research,* vol. 78, pp. 251-261, 2021.

[17] P. Rana, F. Sattari, L. Lefsrud and M. Hendry, "Machine Learning Approach to Enhance Highway Railroad Grade Crossing Safety by Analyzing Crash Data and Identifying Hotspot Crash Locations," *Transportation Research Record,* vol. 2678, no. 7, pp. 1055-1071, 2024.

[18] X. Yang, Q. Li, A. Zhang and Y. Zhan, "Modeling the accident prediction for at-grade highway-rail crossings," *Intelligent Transportation Infrastructure,* vol. 1, 2022.

[19] X. Yin, J. Jin and Z. Zhang, "Interpretable accident prediction at highway-rail grade crossings: a deep learning approach," *Computers & Industrial Engineering,* vol. 207, p. 111337, 2025.

[20] S. Heydari, L. Fu, L. Thakali and L. Joseph, "Benchmarking regions using a heteroskedastic grouped random parameters model with heterogeneity in mean and variance: Applications to grade crossing safety analysis," *Analytic Methods in Accident Research,* vol. 19, pp. 33-48, 2018.

[21] K. Haleem, "Investigating risk factors of traffic casualties at private highway-railroad grade crossings in the United States," *Accident Analysis & Prevention,* vol. 95, no. Pt A, pp. 274-283, 2016.

[22] W. Hao, C. Kamga, X. Yang, J. Ma, E. Thorson, M. Zhong and C. Wu, "Driver injury severity study for truck involved accidents at highway-rail grade crossings in the United States," *Transportation Research Part F Traffic Psychology and Behaviour,* vol. 43, pp. 379-386, 2016.

[23] M. Islam and A. Alogaili, "Uncovering the risks for driver injury severities for truck-trailer and passenger car crashes at highway-railroad crossings," *Transportation Research Interdisciplinary Perspectives,* vol. 31, p. 101451, 2025.

[24] Q. Ren and M. Xu, "Injury severity analysis of highway-rail grade crossing crashes in non-divided two-way traffic scenarios: A random parameters logit model," *Multimodal Transportation,* vol. 3, no. 1, p. 100109, 2024.

[25] S. S. Ahmed, F. Corman and P. C. Anastasopoulos, "Accounting for unobserved heterogeneity and spatial instability in the analysis of crash injury-severity at highway-rail grade crossings: A random parameters with heterogeneity in the means and variances approach," *Analytic Methods in Accident Research,* vol. 37, p. 100250, 2023.

[26] S. Das, X. Kong, S. Lavrenz, L. Wu and M. Jalayer, "Fatal crashes at highway rail grade crossings: A U.S. based study," *International Journal of Transportation Science and Technology,* vol. 11, no. 1, pp. 107-117, 2022.

[27] S. Zhao and A. J. Khattak, "Factors associated with self-reported inattentive driving at highway-rail grade crossings," *Accident Analysis & Prevention,* vol. 109, pp. 113-122, 2017.

[28] G. S. Larue and C. N. Watling, "Prevalence and dynamics of distracted pedestrian behaviour at railway level crossings: Emerging issues," *Accident Analysis & Prevention,* vol. 165, p. 106508, 2022.

[29] C. Wullems, "Towards the adoption of low-cost rail level crossing warning devices in regional areas of Australia: A review of current technologies and reliability issues," *Safety Science,* vol. 49, pp. 1059-1073, 2011.

[30] A. K. Vivek, T. Khan and S. S. Mohapatra, "Safety and associated parameters influencing performance of rail road grade crossings: A critical review of state of the art," *Journal of Safety Research,* vol. 79, pp. 257-272, 2021.

 doi:10.20944/preprints202603.0832.v1

[31] A. Evans and P. Hughes, "Traverses, delays and fatalities at railway level crossings in Great Britain," *Accident Analysis & Prevention*, vol. 129, pp. 66-75, 2019.

[32] S. Lee, T. Chen, N. Sze, T. Mao, Y. Ou, A. Mihăiță and F. Chen, "Analysing driver behaviour and crash frequency at railway level crossings using connected vehicle and GIS data," *Travel Behaviour and Society*, vol. 39, p. 100957, 2025.

[33] M. Naghdi, P. Lautala and A. Erfani, "Assessing Visibility at Highway–Rail Grade Crossings Using High-Resolution LiDAR," *Heliyon*, vol. 10, no. 22, p. e40347, 2024.

[34] S. A. Ibtihal and S. M. Rifaat, "Accident prediction model for Railway Level Crossings (RLCs) in Bangladesh," *KSCE Journal of Civil Engineering*, vol. 30, no. 1, p. 100296, 2025.

[35] USDOT Form 71, "Crossing Inventory Data (Form 71) - Current," 5 March 2026. [Online]. Available: https://data.transportation.gov/Railroads/Crossing-Inventory-Data-Form-71-Current/m2f8-22s6/about_data. [Accessed 5 March 2026].

[36] SimpleMaps, "SimpleMaps.com," Pareto Software, LLC, 2026. [Online]. Available: https://simplemaps.com/data/us-cities. [Accessed 15 February 2026].

[37] M. Bachmann, "GitHub," 2021. [Online]. Available: https://rapidfuzz.github.io/RapidFuzz/. [Accessed 16 February 2026].

[38] OpenAI, "ChatGPT 5.2," [Online]. Available: https://chatgpt.com/. [Accessed 1 February 2026].

[39] Google, "Gemini 3," [Online]. Available: https://gemini.google.com/. [Accessed 1 February 2026].

[40] Anthropic, "Sonnet 4.6," [Online]. Available: https://claude.ai. [Accessed 1 February 2026].

[41] J. Bai and P. Perron, "Computation and Analysis of Multiple Structural Change Models," *Journal of Applied Econometrics*, vol. 18, no. 1, pp. 1-22, 2003.

[42] E. Lehmann and J. P. Romano, Testing Statistical Hypotheses, 4th ed., New York, New York: Springer, 2022, p. 1027.

[43] U.S. Census Bureau, "County Estimated Population 2010-2020," [Online]. Available: https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/. [Accessed 1 February 2026].

[44] FHWA, "Railway Highway Crossing Program Overview," 12 February 2025. [Online]. Available: https://highways.dot.gov/safety/hsip/xings/railway-highway-crossing-program-overview. [Accessed 1 February 2026].

[45] G. J. M. Read, J. A. Cox, A. Hulme, A. Naweed and P. M. Salmon, "What factors influence risk at rail level crossings? A systematic review and synthesis of findings using systems thinking," *Safety Science*, vol. 138, p. 105207, 2021.