FLEETIDY · TECHNICAL METHODOLOGY

The FRED Carrier-Scoring Engine

A complete mathematical specification of how prospective crash burden is estimated, calibrated, and graded for every U.S. motor carrier

---

Document	FRED Scoring — Mathematical White Paper
Audience	Underwriting, actuarial, and executive readers
Population	1,150,553 carriers scored · 1,118,390 graded · 2.1M census
Engine	Gradient-boosted Tweedie burden model with credibility-graded peer ranking
Date	June 2026

§0 Executive Summary

The **FRED score** answers one question for every active for-hire motor carrier in the United States: *relative to peers of the same size, how much crash harm is this carrier likely to generate over the coming year?* The answer is built entirely from federal FMCSA records — fleet size, mileage, roadside inspections, violations, and crashes — and is refreshed weekly across roughly 1.12 million graded carriers.

This paper specifies the engine end to end. We define the quantity being predicted: a carrier's forward twelve-month **severity-weighted crash burden**, in which a fatal crash counts many times a property-damage one. We show how raw event counts are converted into a per-mile **rate** through a mileage exposure model; how thin carrier histories are stabilized with an **Empirical-Bayes** credibility estimator anchored to fleet-size peers; how a **gradient-boosted Tweedie model** learns the relationship between a carrier's observable safety profile and its future burden; how predictions are **calibrated** so that the book as a whole is neither optimistic nor pessimistic; and how the calibrated burden becomes a peer-relative **letter grade** (Excellent … Critical), a 0–100 **score**, and an explicit **fatal-crash probability**. Every constant and formula below is taken directly from the production engine.

> **Three numbers, one model.** For each carrier the engine emits (i) an expected **crash burden** and crash **count** over the next twelve months, (ii) a peer-relative **grade and score** derived from that burden, and (iii) a **fatal-crash probability**. All three share one exposure basis and one set of features, so they tell a mutually consistent story.

§ Contents

§1 What the Score Estimates

Underwriters do not ultimately care how many violations a truck has — they care about the **cost of future crashes**. FRED therefore predicts a forward-looking, harm-weighted quantity rather than a backward-looking tally. For carrier i, define the target

$$B_i \;=\; \text{total severity-weighted crash burden generated by carrier } i \text{ over the next 12 months.}$$

A single crash contributes a weight that reflects its severity (§4): a fatality weighs far more than a fender-bender. B_i is a non-negative quantity that is **exactly zero for the large majority of carriers** (most have no crash in a given year) and continuously positive for the rest. That two-part structure — a spike of probability mass at zero plus a continuous positive tail — is the defining feature of the data and dictates the model family we use (§7).

Because a 3-truck owner-operator and a 5,000-truck fleet are not comparable on raw counts, the engine works in **rates per mile driven** and then expresses every carrier **relative to its own size peers**. The published grade answers "how does this carrier compare to fleets like it," while the underlying expected burden and crash count answer "how much, in absolute terms."

---

§2 Data and Notation

All inputs are federal FMCSA datasets, refreshed weekly: the **census** (fleet size, mileage, domicile, authority, safety rating), **roadside inspections**, **violations**, and **crashes**. The engine scores every carrier with a positive reported power-unit count.

The model is trained on a **rolling year-over-year window**. Let C be the most recent crash-mature date (we discard the last 45 days, since crash reports arrive with a lag). We form two consecutive **half-open** windows of exactly 365 days each:

$$\underbrace{[\,C-730,\;C-365\,)}_{\text{feature window } W_{\text{fit}}} \;\longrightarrow\; \underbrace{[\,C-365,\;C\,)}_{\text{outcome window } W_{\text{out}}} \quad .$$

Each window includes its left endpoint and excludes its right, so the two windows partition the 730 days before C: no day is double-counted and none is dropped.

The model learns the map from a carrier's safety profile in W_{fit} to its *realized* burden in the following year W_{out}. To score live carriers we then feed each carrier's **most recent** twelve months of features (the W_{out} window) into the fitted model to predict the *forward* twelve months. This out-of-time design is what makes the score genuinely predictive rather than a restatement of past events.

Table 2.1 — Core notation used throughout.

Symbol	Meaning
E_i	Exposure of carrier i, in units of 100,000 miles per year (§3)
N_i	Crash count in a twelve-month window
B_i	Severity-weighted crash burden in a twelve-month window (§4)
$b(i)$	Fleet-size band of carrier i: small / medium / large / xlarge (§3)
$\mathbf{x}_i$	Feature vector of carrier i (§6)
$\hat{B}_i,\ \hat{N}_i$	Model-predicted forward burden and crash count

§3 Exposure: Turning Fleets into Miles

Crash risk scales with how much a carrier drives, so **miles** — not trucks — is the natural denominator. We measure exposure in units of 100,000 miles per year,

$$E_i \;=\; \frac{\text{annual miles driven by carrier } i}{10^5}.$$

Reported mileage is sometimes missing or implausible, so the engine applies a reliability test before trusting it. A carrier's mileage is deemed **reliable** when miles are positive, the implied miles-per-power-unit lies in a sane band, and FMCSA's own outlier flags are clear:

$$\text{reliable}_i \;=\; \big[\, m_i > 0 \,\big] \wedge \Big[\, 1{,}000 \le \tfrac{m_i}{u_i} \le 300{,}000 \,\Big] \wedge \big[\, \neg\,\text{outlier}_i \,\big],$$

where m_i is reported annual miles and u_i is power units. When mileage is unreliable, exposure is **imputed** from the carrier's fleet size using the median miles-per-truck of its size band $b(i)$:

$$E_i \;=\; \begin{cases} m_i / 10^5, & \text{reliable}_i, \\ \bar{m}_{b(i)} \cdot u_i / 10^5, & \text{otherwise}, \end{cases} \qquad E_i \leftarrow \min\!\big(\max(E_i,\, 10^{-6}),\, 30{,}000\big).$$

Size bands are fixed by power-unit count and are the unit of peer comparison everywhere downstream:

Table 3.1 — Fleet-size bands.

Band	Power units u	Typical carrier
small	$u \le 5$	Owner-operator / micro-fleet
medium	$6 \le u \le 20$	Small regional
large	$21 \le u \le 100$	Mid-size fleet
xlarge	$u > 100$	Large carrier

Finally, two **data-quality guards** protect the book from corrupt federal records. A carrier whose reported fleet exceeds 50,000 power units (the real-world maximum is ~20,000), or which claims a very large fleet with no usable mileage to corroborate it, is treated as having no trustworthy exposure and is left **ungraded** rather than handed a fabricated baseline.

---

§4 Severity Weighting: Measuring Harm

Not all crashes are equal. The burden of a single crash c is a weight that rises with its consequences — counting casualties, not merely flagging them:

$$w_c \;=\; 1 \;+\; 12 \cdot \min(\text{fatalities}_c,\, 3) \;+\; 4 \cdot \min(\text{injuries}_c,\, 5) \;+\; 3 \cdot \mathbb{1}[\text{hazmat released}].$$

A baseline tow-away crash weighs 1; a one-injury crash weighs 5; a one-fatality crash weighs at least 13 — and a crash that kills two people weighs more than one that kills one. Fatality and injury terms **stack**: a fatal crash with two additional injuries weighs $1 + 12 + 8 = 21$. The caps (three fatalities, five injuries) keep a single catastrophic record from dominating the training target. A carrier's burden over a window is the sum over its crashes,

$$B_i \;=\; \sum_{c \,\in\, \text{crashes}(i)} w_c,$$

and the crash **count** is simply $N_i = |\text{crashes}(i)|$. These weights are a deliberately **compressed harm index**, chosen for statistical stability rather than cost proportionality: DOT's value of a statistical life implies a fatal-to-property-damage cost ratio in the *hundreds*, and a cost-proportional target would be dominated by rare fatal events and untrainable. Tail risk is instead carried by the dedicated fatal-crash model of §11, and Appendix §B shows that carrier rankings are nearly invariant to the choice of fatal weight (8 / 12 / 20).

EXAMPLE 4.1 — BURDEN OF A SMALL FLEET

A 12-truck carrier had three crashes last year: a crash with two fatalities and one injury, a hazmat-release crash, and a tow-away. Its burden is

$$B \;=\; \underbrace{(1 + 2 \cdot 12 + 1 \cdot 4)}_{\text{2 fatalities + 1 injury}} \;+\; \underbrace{(1 + 3)}_{\text{hazmat}} \;+\; \underbrace{1}_{\text{tow}} \;=\; 34, \qquad N = 3.$$

The fatal and injury contributions of the first crash stack ($1 + 24 + 4 = 29$); under indicator weights it would have counted the same as a single-fatality crash. A peer with three tow-aways has the same count but only $B = 3$ — the model treats the first carrier as by far the larger risk.

§5 Empirical-Bayes Peer Relativities

A carrier's own history is the strongest signal we have — but for a small fleet it is also the noisiest. One crash for a 3-truck operator could be terrible luck or a genuine pattern; the raw rate alone cannot tell us which. The classical actuarial answer is **credibility**: blend the carrier's own experience with the experience of its peers, weighting the carrier more as its data accumulates. We implement this with an **Empirical-Bayes Gamma–Poisson** model, fit separately within each size band.

Within a band, suppose each carrier's latent crash rate λ is drawn from a Gamma prior, and crashes are Poisson given the rate:

$$\lambda \sim \text{Gamma}(\alpha, \beta), \qquad N \mid \lambda, E \sim \text{Poisson}(\lambda E).$$

The prior is estimated from the band's pooled data with the standard **Bühlmann–Straub** estimators for Poisson exposure data. Write $r_k = N_k/E_k$ for carrier rates and $E_\bullet = \sum_k E_k$ for total band exposure. The grand mean is the exposure-weighted rate — total events over total exposure, not a mean of carrier rates:

$$\hat{\mu} \;=\; \frac{\sum_k N_k}{\sum_k E_k}.$$

A subtlety matters here. The raw variance of observed rates, $\sum_k \pi_k (r_k - \hat{\mu})^2$, is **not** the between-carrier variance: conditional on a carrier's true rate λ, the observed rate still fluctuates with Poisson sampling noise of variance λ/E_k, so the raw quantity estimates $\text{Var}(\lambda) + \mathbb{E}[\lambda/E_k]$. Using it directly would overstate the prior variance, understate β, and hand thin-data carriers too much credibility. Bühlmann–Straub removes the sampling noise explicitly: for Poisson data the process variance is $\hat{s}^2 = \hat{\mu}$, and the variance of hypothetical means (the true between-carrier variance) is

$$\hat{a} \;=\; \frac{\sum_k E_k\,(r_k - \hat{\mu})^2 \;-\; (n_b - 1)\,\hat{s}^2}{E_\bullet \;-\; \sum_k E_k^2 / E_\bullet},$$

where n_b is the number of carriers in the band. If the numerator is non-positive the band shows no detectable between-carrier heterogeneity, and $\hat{a}$ is floored at a small ε so that credibility simply falls to zero rather than dividing by zero. No tail trimming is applied to the estimation sample; robustness to corrupt records comes from the §3 data-quality guards and the relativity clip, not from biasing $\hat{a}$. The Gamma prior parameters follow:

$$\beta = \frac{\hat{\mu}}{\hat{a}}, \qquad \alpha = \hat{\mu}\,\beta,$$

and this β *is* the Bühlmann credibility constant K of classical credibility theory.

Gamma–Poisson conjugacy then gives the posterior mean rate for a carrier with N events over exposure E in closed form, and we report it as a **relativity** against the band mean $\hat{\mu}$:

$$\boxed{\rho \;=\; \frac{1}{\hat{\mu}} \cdot \frac{\alpha + N}{\beta + E}} \qquad (\rho = 1 \text{ means “exactly typical for your size.”})$$

This is exactly a credibility blend. Rewriting with credibility weight $Z = E/(E+\beta)$,

$$\frac{\alpha + N}{\beta + E} \;=\; Z \cdot \underbrace{\frac{N}{E}}_{\text{own rate}} \;+\; (1 - Z) \cdot \underbrace{\hat{\mu}}_{\text{peer mean}} \;,$$

so a carrier with little exposure is pulled toward its peers ($Z \to 0$), while a carrier with a long track record stands on its own experience ($Z \to 1$). The prior dispersion β sets the "speed" of that transition and is learned from data, not assumed.

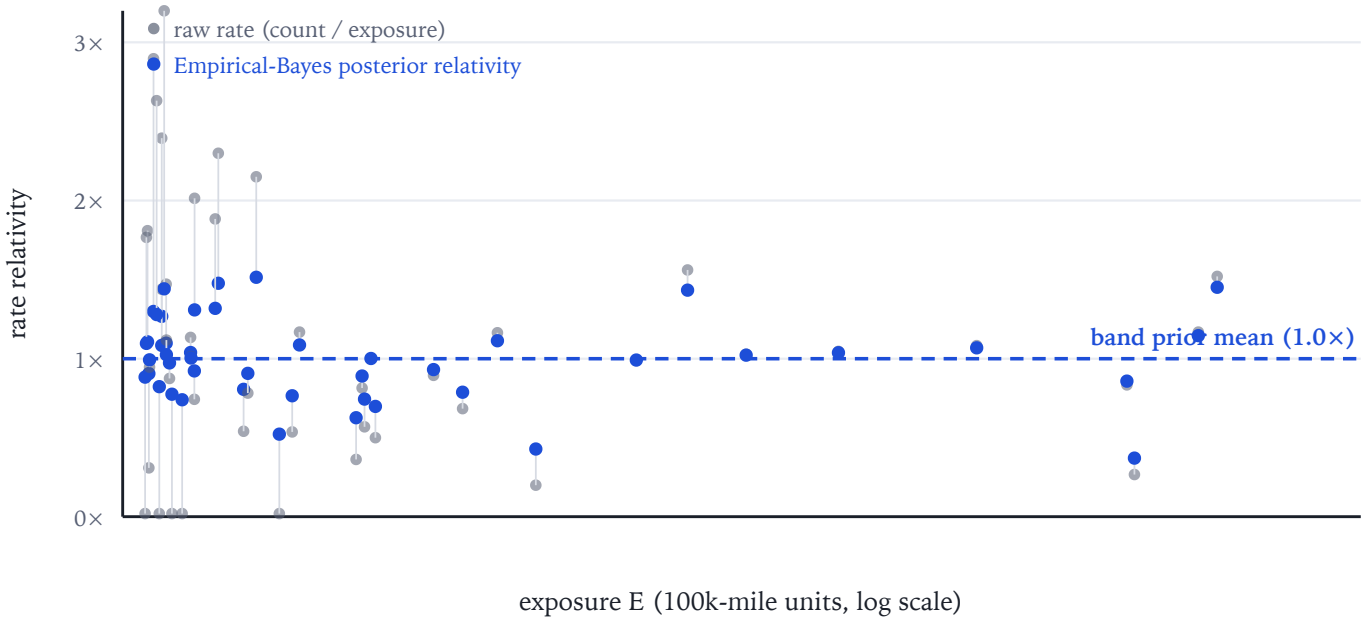

Figure 5.1 — Empirical-Bayes shrinkage. Raw per-mile rates (grey) are noisy at low exposure; the posterior relativities (blue) are pulled toward the band mean by an amount that shrinks as exposure grows. High-exposure carriers barely move.

The same Gamma–Poisson machinery produces four relativities that become model inputs — for **crashes** (exposure = miles) and, using inspections as the exposure base, for **behavioral**, **equipment**, and **severe** violations. Each relativity is clipped to $[0.01, 100]$ for numerical safety.

---

§6 The Feature Vector

Each carrier is summarized by twenty covariates $\mathbf{x}_i$ built from its trailing-year record. They fall into four interpretable groups. Counts enter on a $\log(1+\cdot)$ scale (diminishing marginal signal), rates enter as fractions, and the Empirical-Bayes relativities enter as logs so that "twice typical" and "half typical" are symmetric.

Table 6.1 — The 20 features, by group.

Group	Features
Size	band indicators for medium, large, xlarge (small is the reference)
Risk relativities (§5)	$\log \rho_{\text{crash}}$, $\log \rho_{\text{behavioral}}$, $\log \rho_{\text{equipment}}$, $\log \rho_{\text{severe}}$
Inspection record	$\log(1 + \text{inspections})$, no-inspection flag, driver out-of-service rate, vehicle out-of-service rate, inspection intensity $\log(1 + \text{inspections}/E)$
Violation profile	$\log(1 + \text{unsafe})$, $\log(1 + \text{hours-of-service})$, $\log(1 + \text{maintenance})$, speeding rate, reckless-driving flag
Operating profile	years in business (capped at 30, scaled), interstate indicator, high-utilization flag (> 200k miles/truck)

The **inspection-intensity** feature deserves a word. Roadside inspection is not random sampling — enforcement targets carriers that look risky, so a per-inspection violation rate understates the gap between good and bad fleets (a heavily targeted carrier faces tougher scrutiny per stop). Including $\log(1 + \text{inspections}/E)$ lets the trees condition the per-inspection relativities on how intensively the carrier is selected for inspection, absorbing that selection effect instead of letting it bias the violation signals.

Crucially, federal risk scores that themselves depend on crashes (FMCSA BASIC percentiles, prior FRED outputs) are **excluded** from $\mathbf{x}$: admitting them would leak the outcome into the features and inflate apparent accuracy. The model learns only from primary, behaviorally meaningful signals.

---

§7 The Burden Model: Tweedie + Gradient Boosting

The target B_i has the awkward shape noted in §1 — a point mass at zero plus a continuous positive tail — so neither a pure count model nor an ordinary regression fits it. The natural distribution for exactly this "zero-inflated continuous" structure is the **Tweedie** family, a compound Poisson–Gamma law in which a Poisson number of crashes each carry a Gamma-distributed severity. Its variance is a power of its mean,

$$\operatorname{Var}(B) \;=\; \phi\, \mathbb{E}[B]^{\,p}, \qquad 1 < p < 2,$$

and for any p in that range the model places explicit probability on $B = 0$ while modelling the positive tail smoothly. This single choice is what lets one model speak to the whole population at once.

The variance power p is **estimated, not assumed**. Each refresh runs a profile search over the grid $p \in \{1.1, 1.2, \ldots, 1.9\}$: the burden head is retrained at every candidate on a carrier-disjoint 80% split and evaluated on the held-out 20%. Because Tweedie deviance changes scale with p, raw (or even null-normalized) deviance is not comparable across candidates; selection therefore uses a p-free out-of-time criterion — the exposure-weighted normalized Gini of the holdout ranking (§13) — with the holdout deviance at each candidate logged alongside. The selected p^*, and the metric's sensitivity at $p^* \pm 0.1$, are

persisted as logged production constants on every run; prediction *levels* are pinned downstream by the §9 calibration regardless of p.

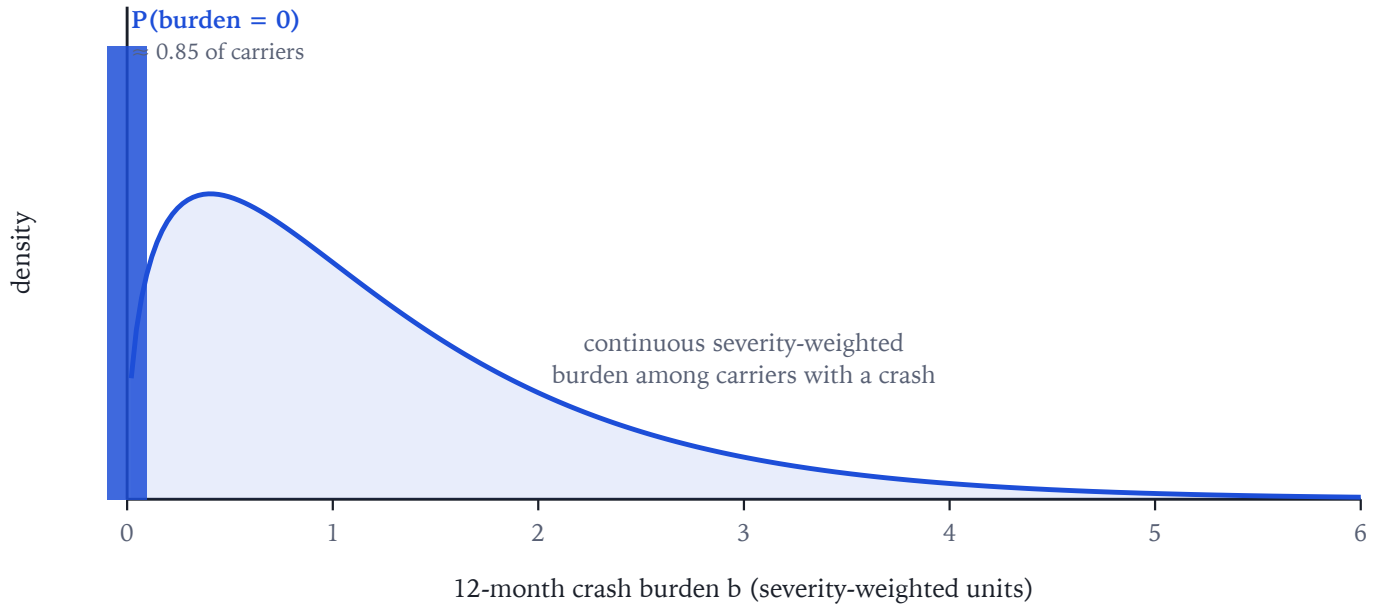

Figure 7.1 — The Tweedie law matches the data. A large point mass at zero (most carriers have no crash) sits alongside a continuous, right-skewed distribution of positive burden for those that do. A Gaussian or pure-count model cannot represent both.

We do not assume the relationship between features and burden is linear. Instead we learn it with **gradient-boosted regression trees** (XGBoost) under the Tweedie deviance loss. The model is an additive ensemble of shallow trees $f = \sum_t f_t$, grown stage-wise so that each new tree corrects the errors of those before it. The predicted forward burden is

$$\mathbb{E}[B_i \mid \mathbf{x}_i] \;=\; \exp\Big(\underbrace{\log E_i}_{\text{base margin}} \;+\; f(\mathbf{x}_i) \Big)$$

The term $\log E_i$ is a fixed **base margin** (an offset), not a fitted coefficient. Its effect is decisive: because it enters the log-mean with coefficient one, the trees f are forced to predict a burden **rate per mile**, and total burden scales automatically with exposure. A carrier that drives twice as far, all else equal, is predicted to generate twice the burden — without the model having to relearn that fact.

Trees are well suited to this problem: they capture interactions (e.g. a high out-of-service rate matters more for a high-mileage long-haul fleet) and non-linear thresholds without hand-specification, and — unlike a linear model — they need no feature standardization. Over-fitting is controlled by shallow depth, slow learning, strong leaf-size regularization, and row/column subsampling:

Table 7.1 — Burden-model hyperparameters (production).

Parameter	Value	Role
objective	Tweedie, p^* selected per refresh (currently 1.1)	zero-inflated continuous loss
max tree depth	4	limits interaction order
learning rate	0.05	slow, stable boosting
boosting rounds	400	ensemble size
min child weight	30	no leaf on <30-carrier evidence
L_1, L_2 penalties	0.1, 1.0	shrink leaf weights
row / column subsample	0.8 / 0.8	de-correlate trees

§8 The Frequency Model

Alongside burden we fit a second, parallel ensemble for crash **count**, using the same features and the same log-exposure base margin but a **Poisson** objective:

$$\mathbb{E}[N_i \mid \mathbf{x}_i] \;=\; \exp\big(\log E_i + g(\mathbf{x}_i)\big).$$

The frequency head drives the user-facing "expected crashes in the next 12 months," the typical-fleet baseline shown next to it, and — as we will see in §9 — the anchor for calibrating the book. Separating frequency from burden lets the product say both "how often" and "how badly," and their ratio $\hat{B}_i/\hat{N}_i$ is an implied **severity** per crash.

§9 Calibration to Experience

A model can rank carriers correctly yet still be systematically high or low in **level** — predicting, say, 10% more crashes than actually occur. For pricing, level matters. We therefore **calibrate within each size band** so that predicted totals match observed totals (an observed-to-expected ratio of one). For band b, the count and burden scaling factors are

$$\kappa_b^N \;=\; \frac{\sum_{i\in b} N_i}{\sum_{i\in b} \hat{N}_i}, \qquad \kappa_b^B \;=\; \frac{\sum_{i\in b} B_i}{\sum_{i\in b} \hat{B}_i},$$

and every carrier's prediction is rescaled, $\hat{N}_i \leftarrow \kappa^N_{b(i)}\hat{N}_i$ and $\hat{B}_i \leftarrow \kappa^B_{b(i)}\hat{B}_i$. Because burden is modelled directly (not derived from counts), it gets its **own** calibration factor against observed burden — counts and burden are anchored independently. Anchoring on the most recent observed year, rather than on the

training book, removes two biases that would otherwise inflate large-fleet risk: the training set over-represents active carriers, and imputed mileage distorts exposure.

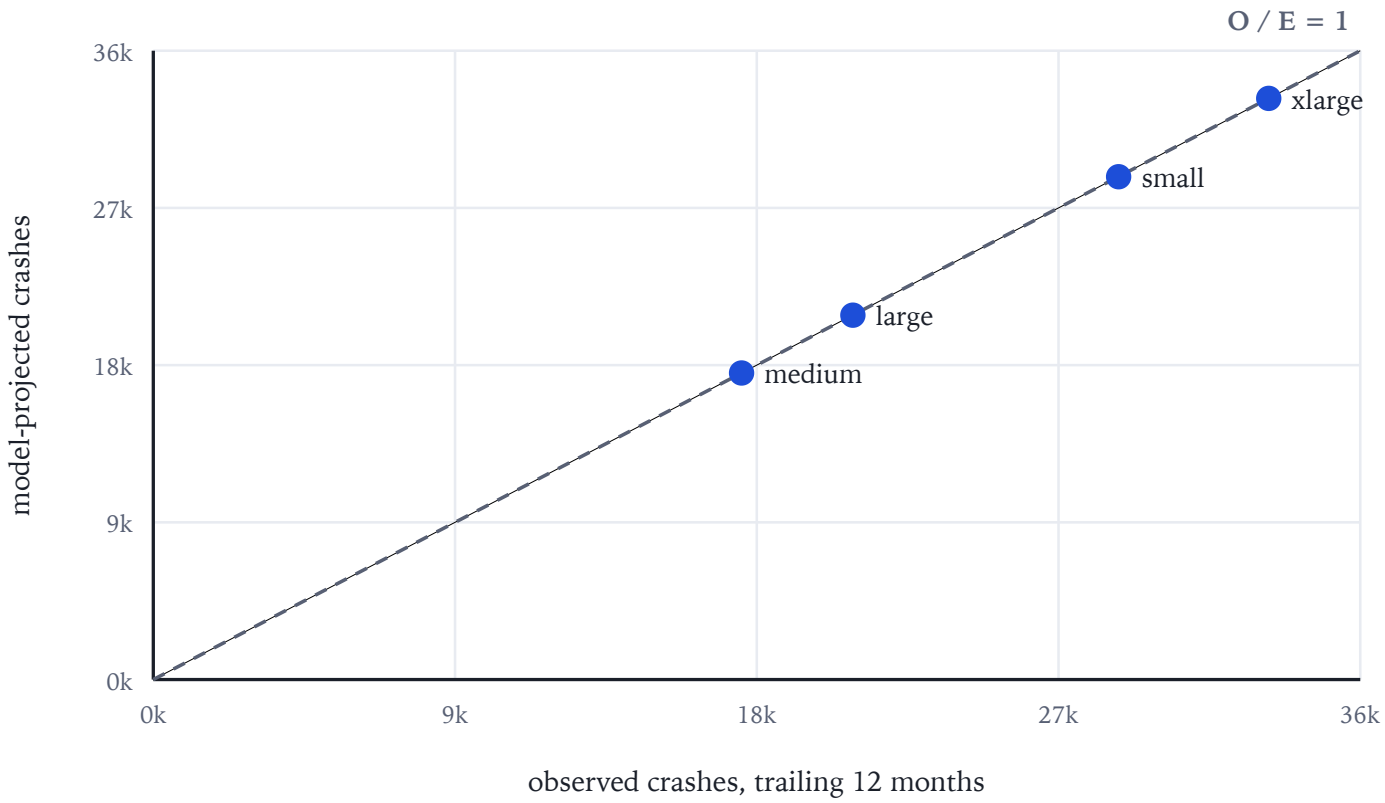

Figure 9.1 — After per-band calibration, model-projected crashes equal observed crashes in every size band (points lie on the 45° line). Calibration fixes level; the model's job is to get the ordering right (Figure 13.1).

---

§10 From Burden to Grade

The published grade is **peer-relative**: it measures a carrier against fleets of its own size, so a careful owner-operator and a careful national fleet can both earn top marks. Three steps turn calibrated burden into a grade.

Step 1 — within-band relativity

Convert burden to a per-mile rate and divide by the **exposure-weighted mean** rate of the carrier's band — the same kind of center the Empirical-Bayes machinery uses in §5:

$$\tilde{r}_i = \frac{\hat{B}_i / E_i}{\bar{R}_{b(i)}}, \qquad \bar{R}_b = \frac{\sum_{j \in b} \hat{B}_j}{\sum_{j \in b} E_j}.$$

Here $\tilde{r}_i = 1$ means the carrier matches the **band-average fleet** — the burden rate of the band's total book — and $\tilde{r}_i = 2$ is twice that average. One consequence worth stating plainly: burden is right-skewed, so the *median* carrier in a band sits well below the band average (typically around $\tilde{r} \approx 0.5$–0.6). The same center, 1.0 = band average, is also the Step-2 shrinkage target and the basis of every “average fleet of your size” baseline shown in the product.

Step 2 — credibility shrinkage

A thin-data carrier should not be branded by a single noisy year. We shrink its relativity toward the band average (1.0) using a Bühlmann credibility weight. The shrinkage constant is **not** the crash-count β_b of §5: burden carries severity variance on top of count variance, so per unit of exposure a burden relativity deserves *less* credibility than a count relativity. The constant comes from the compound-Poisson structure of burden. With crashes at rate λ and i.i.d. per-crash weights w,

$$\operatorname{Var}(B \mid \lambda, E) \;=\; \lambda E\,\mathbb{E}[w^2] \;\Longrightarrow\; \hat{s}_B^2 \;=\; \hat{\mu}_B\,\frac{\mathbb{E}[w^2]}{\mathbb{E}[w]}, \qquad \hat{\mu}_B = \frac{\sum_k B_k}{\sum_k E_k},$$

where $\mathbb{E}[w]$ and $\mathbb{E}[w^2]$ are estimated per band from the realized crash-weight distribution in the training window. The between-carrier burden variance $\hat{a}_B$ uses the same Bühlmann–Straub estimator as §5 (substituting B for N and $\hat{s}_B^2$ for $\hat{s}^2$), giving

$$K_{B,b} \;=\; \frac{\hat{s}_B^2}{\hat{a}_B}, \qquad Z_i \;=\; \frac{E_i}{E_i + K_{B,b(i)}}, \qquad s_i \;=\; Z_i\,\tilde{r}_i + (1 - Z_i)\cdot 1.$$

K_B exceeds the count β in every band — correct and expected, since the severity term $\mathbb{E}[w^2]/\mathbb{E}[w]$ multiplies the process variance. The per-band K_B values are logged as production constants on every refresh (Appendix §A).

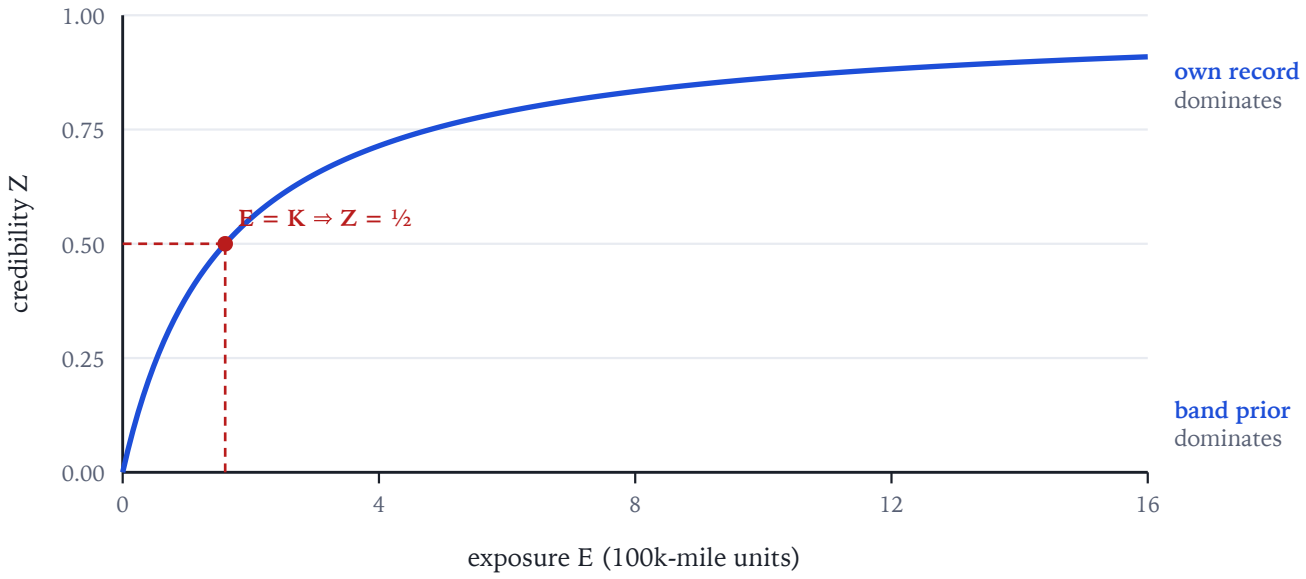

Figure 10.1 — The credibility weight $Z = E/(E+K)$. A carrier is graded on its own record once its exposure exceeds the band's burden credibility constant $K = K_{B,b}$; below that, its grade leans on its peers. The curve is steep where it matters and flat where data is ample.

Step 3 — within-band percentile, grade, and score

Rank the shrunken relativities s_i **within the carrier's own size band**, against the other gradeable carriers of that band only, and read off each carrier's percentile $p_i \in [0, 1]$ (0 = safest, 1 = riskiest). Ranking within band is what makes the grade mean what the product says it means — "compared to fleets like yours": band distributions of s_i have different spreads (small carriers are shrunk harder), so a single population-wide ranking would squeeze small carriers into the middle grades and hand both tails to large carriers.

Exact ties — common among zero-history carriers — share their average rank, so identical records always receive identical grades. The letter grade is a fixed partition of the within-band percentile, and the 0–100 score is its complement:

$$\text{score}_i \;=\; 100\,(1 - p_i).$$

Table 10.1 — Grade cut-points on the within-band percentile p.

Grade	Percentile band	Reading
Excellent	$p \le 0.08$	safest 8% of peers
Strong	$0.08 < p \le 0.25$	well above typical
Satisfactory	$0.25 < p \le 0.70$	around typical
Marginal	$0.70 < p \le 0.87$	below typical
Poor	$0.87 < p \le 0.95$	elevated risk
Critical	$p > 0.95$	riskiest 5%

Carriers whose exposure is too thin to credit (a **provisional** tier) are **capped** at Satisfactory: a carrier cannot earn a top grade without a track record, but it can still be flagged as risky. As exposure accrues the cap lifts automatically. Each carrier also carries a plain-English **confidence tier** — High, Moderate, Low, or Prior-only — set by its credibility Z_i, so a reader always knows how much of the grade is the carrier's own record versus its peer prior.

Ordering matters and is explicit. Percentiles are computed first, on the gradeable population; the provisional cap and the FMCSA-Unsatisfactory override (§12) are applied *after* percentiling. The post-override distribution therefore deviates from the nominal cut-points — Critical exceeds 5% (forced overrides) and Satisfactory exceeds 45% (capped provisionals) — and the engine logs both the pre-override and post-override distributions on every run; the current production figures appear in §13.

---

§11 The Fatal-Crash Model

Fatal crashes are rare enough that the burden weighting alone does not fully characterize tail risk, so the engine fits a dedicated **fatal-frequency** model. Because fatal events are sparse, a parsimonious **Poisson generalized linear model** on the same features and log-exposure offset is more stable here than a deep ensemble:

$$\mathbb{E}[N_i^{\text{fatal}}] \;=\; \exp\!\big(\log E_i + \mathbf{x}_i^{\top}\boldsymbol{\beta}\big),$$

fit by iteratively reweighted least squares and calibrated per band exactly as in §9. Treating the calibrated fatal intensity as a Poisson mean λ_i, the probability of **at least one** fatal crash in the next year is

$$\Pr(\ge 1 \text{ fatal crash})_i \;=\; 1 - e^{-\lambda_i}.$$

This converts an abstract intensity into a number an underwriter can act on directly, and is reported alongside a typical-fleet fatal baseline for context.

§12 Eligibility and Overrides

A small set of deterministic rules sits on top of the statistical model, encoding facts that should override any learned prediction. They are applied **after** the §10 percentiling (see "Ordering matters" there):

Table 12.1 — Override and eligibility rules.

Condition	Action
FMCSA **Unsatisfactory** safety rating	Grade forced to **Critical**, score 0
No active operating authority, or no reported power units	Left **ungraded** (N/A)
Implausible fleet size / corrupt exposure (§3)	Left **ungraded** (N/A)
Provisional (thin) exposure	Grade **capped at Satisfactory**

These rules are intentionally few. The philosophy is that the model should do the work; overrides exist only for regulatory facts (an adverse federal rating) and data-integrity protection (un-scoreable records), never as hand-tuned score nudges.

§13 Validation and Live Results

Calibration (§9) guarantees the right *level* by construction, so validation focuses on **discrimination** — does the model concentrate future burden in the carriers it flags? The primary metric is the exposure-weighted **ordered-Lorenz Gini**. Rank carriers from safest to riskiest by predicted burden rate, plot cumulative share of exposure against cumulative share of *realized* burden, and measure how far that curve bows below the diagonal. We report the **normalized** Gini — the model's Gini divided by an oracle that knows the realized order — so 1.0 is perfection and 0 is random.

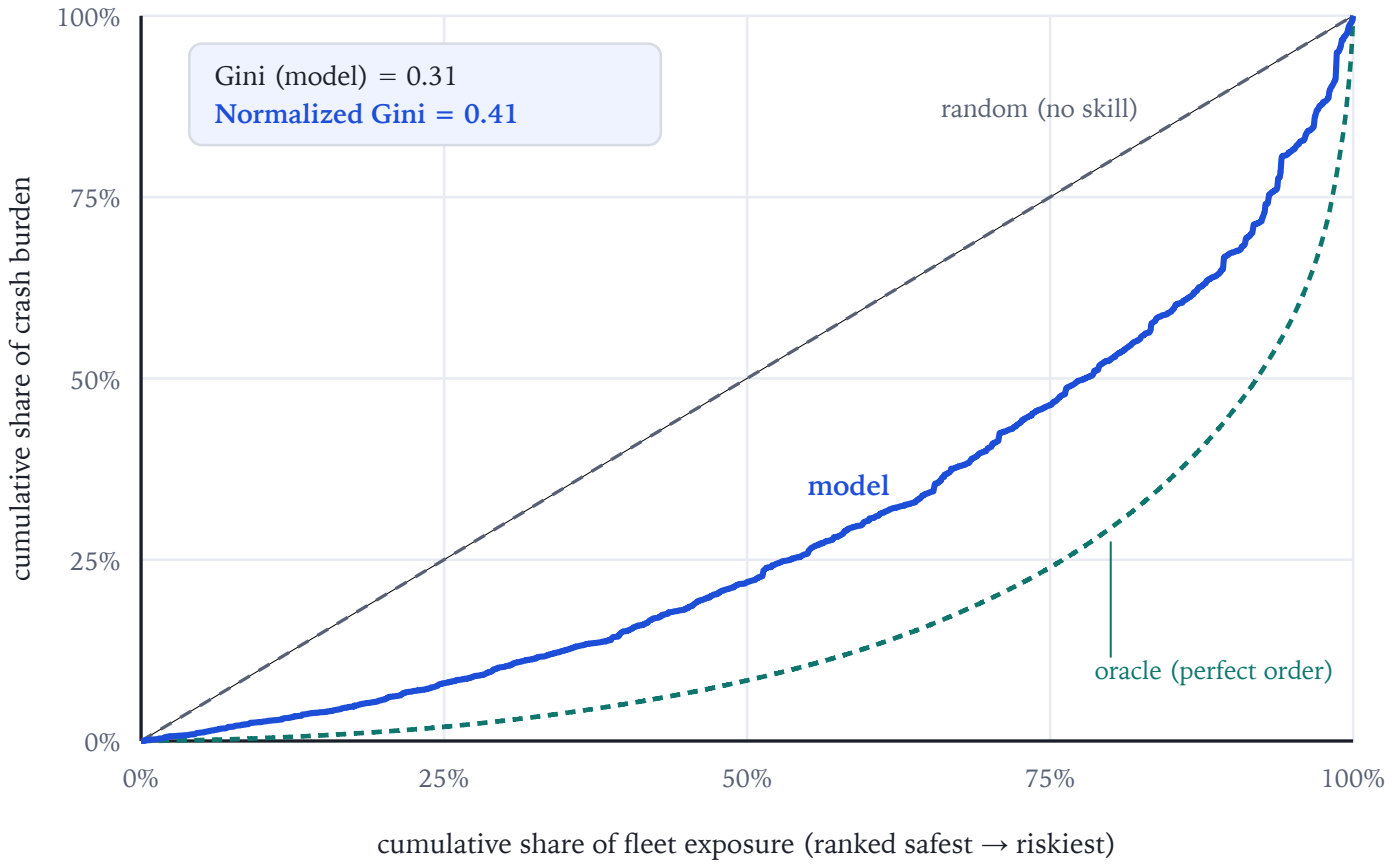

Figure 13.1 — Ordered-Lorenz (Gini) curve on out-of-time data. The model (solid) bows well below random ranking (diagonal) toward the oracle that knows the realized order (dashed). On the carrier-disjoint holdout the engine reaches a normalized Gini of 0.41 (raw model Gini 0.31).

Every refresh runs the full validation suite on a **carrier-disjoint, out-of-time holdout**: one carrier in five (by DOT number) is excluded from fitting, and every figure below is measured on those held-out carriers' realized next-year burden. On the June 2026 refresh (holdout $n = 56{,}269$) the burden model attains an overall normalized Gini of **0.41**; the pre-fix engine, measured under the identical protocol, reached 0.40.

Book-level Gini, however, can be flattered by the model merely learning fleet size, so the suite reports discrimination **within each band**, alongside the realized burden share of the riskiest-ranked 10% of carriers (top-decile share) and a per-band level check whose calibration factors come from the training split only — so, unlike §9's in-sample anchoring, this O/E is *not* one by construction:

Table 13.1 — Per-band out-of-time validation, June 2026 refresh (carrier-disjoint holdout).

Band	Holdout n	Normalized Gini	Top-decile share	O/E (burden)	O/E (count)	Grades monotone
small	43,072	0.27	16.9%	1.06	1.03	yes
medium	9,171	0.25	12.9%	1.00	0.99	yes
large	3,334	0.33	13.0%	1.00	0.99	yes
xlarge	692	0.61	10.1%	0.97	0.99	yes

The last column is the **grade-monotonicity gate**: held-out carriers are pushed through the full §10 pipeline with train-derived constants, and their realized burden rates must increase across grade buckets in every band. In the small band, for example, realized burden rates run 0.18 (Excellent) → 0.20 → 0.28 → 0.38 → 0.44 → 1.25 (Critical) per 100k miles. A refresh that fails this gate aborts before writing — the combined input-EB and output-shrinkage cannot drift out of order silently. As elsewhere: calibration fixes *level* by construction, so validation tests *ordering* — and the holdout O/E is the one level check that has to be earned.

Run against the full federal population in June 2026, the engine scores **1,150,553** carriers, of which **1,118,390** receive a letter grade (the remainder are ineligible — no authority or un-scoreable exposure, §12). Per §10, percentiles are within-band and overrides land afterwards: the pre-override distribution tracks the nominal cut-points in every band (Critical 5.0%, Satisfactory 46.5% — exact-tie blocks share one grade, so shares deviate slightly from the nominal partition), while the realized post-override distribution has **Critical at 5.02%** (FMCSA-Unsatisfactory forces 204 extra carriers in) and **Satisfactory at 49.2%** (the provisional cap moves about 30,000 thin-data carriers down from Strong). Both distributions are logged on every run. The graded distribution below is middle-weighted: most carriers cluster around their band's middle grades, with thin tails of Excellent and Critical, exactly as a well-behaved peer-relative score should.

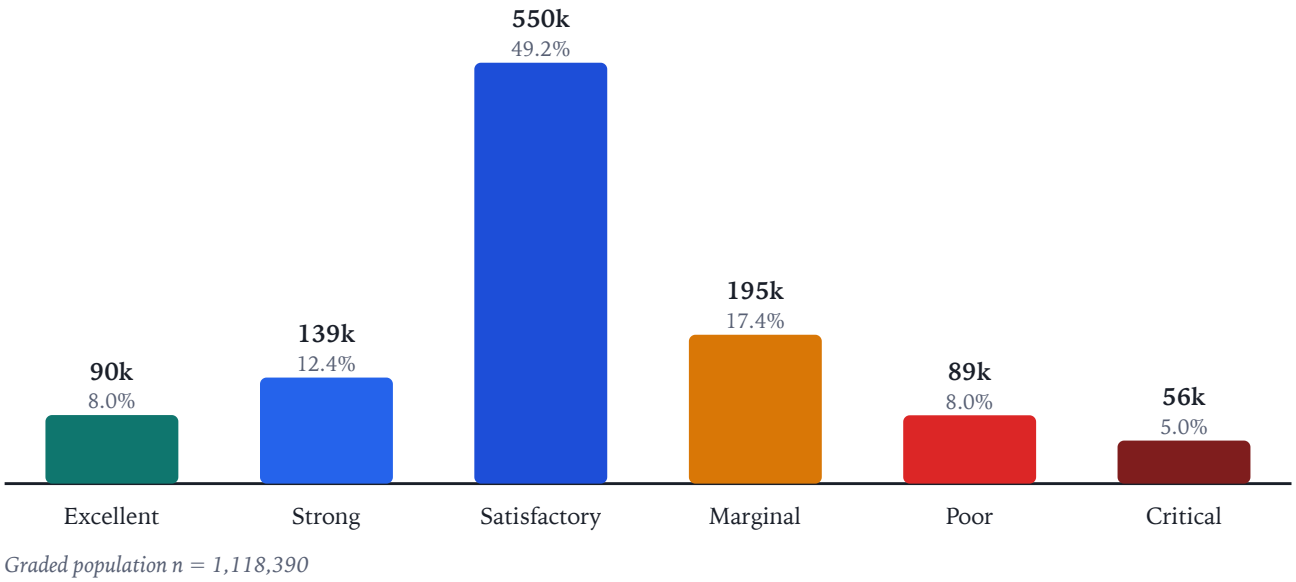

Figure 13.2 — Live grade distribution, June 2026 production run (1,118,390 graded of 1,150,553 scored), post-override. The shape is deliberately middle-heavy: grades are within-band percentiles of peer-relative risk, not absolute pass/fail thresholds.

§14 A Carrier, End to End

WORKED EXAMPLE — A 15-TRUCK REGIONAL FLEET

Inputs. 15 power units (band: medium), 1.65M reliable miles last year, 16 inspections, an elevated vehicle out-of-service rate, two crashes (one with a single injury, one tow-away), no fatal crashes.

Exposure. $E = 1{,}650{,}000/10^5 = 16.5$ (100k-mile units) — reliable, so used directly.

Observed burden (§4). $B = (1+4)+1 = 6$ over $N = 2$ crashes.

Crash relativity (§5). With the medium-band Bühlmann–Straub prior, two crashes over 16.5 exposure units yield a posterior crash relativity of $\rho_{\text{crash}} \approx 1.3$ — observed crash frequency about 30% above same-size peers, credibly so given solid exposure.

Model prediction (§7–9). Feeding $\mathbf{x}$ (band, the four log-relativities, inspection and violation profile, inspection intensity, …) through the boosted Tweedie head and applying the medium-band calibration factor yields a forward burden $\hat{B} \approx 3.5$ and forward count $\hat{N} \approx 1.1$ crashes — against a band-average baseline of about 0.9 crashes for this exposure.

Grade (§10). The carrier's calibrated burden rate is $3.5/16.5 \approx 0.21$ per 100k miles, against a medium-band exposure-weighted average of 0.19: $\tilde{r} \approx 1.13$. With $K_B \approx 13.8$ for the medium band, $Z = 16.5/(16.5+13.8) \approx 0.54$, so the shrunken relativity is $s \approx 0.54(1.13) + 0.46(1.0) = 1.07$. Because burden is concentrated in a risky minority, most of the band sits well below the average, and $s = 1.07$ lands near the 80th percentile *of the medium band* — a **Marginal** grade, score ≈ 20, confidence tier **High**.

Fatal risk (§11). The fatal head returns $\lambda \approx 0.03$, so $\Pr(\geq 1 \text{ fatal}) = 1 - e^{-0.03} \approx 3\%$ over the year.

Story told. "A medium fleet driving its expected miles. Its *observed* crash frequency runs about 30% above same-size peers ($\rho_{\text{crash}} \approx 1.3$); its *predicted* forward burden runs about 13% above the medium-band average fleet ($\tilde{r} \approx 1.13$), which places it near the 80th percentile of its band. Enough history to trust the read. Expect roughly 1.1 crashes — versus 0.9 for the average fleet this active — and a 3% chance of a fatal one in the coming year."

§A Symbols and Constants

Table A.1 — Production constants.

Constant	Value	Section
Crash maturity lag	45 days	§2
Exposure unit	100,000 mi/yr	§3
Reliable miles-per-truck band	[1,000 , 300,000]	§3
Exposure ceiling	30,000 (3B mi/yr)	§3
Corrupt fleet-size guard	> 50,000 power units	§3
Severity weights (per fatality / per injury / hazmat)	12 / 4 / 3	§4
Severity caps (fatalities / injuries)	3 / 5	§4
EB prior estimator	Bühlmann–Straub, no trimming	§5
Relativity clip	[0.01 , 100]	§5
Estimation exposure floor	$E \geq 10^{-3}$ (excludes §3 sentinels)	§5
GBM features	20	§6
Tweedie variance power p	selected per refresh from {1.1 … 1.9}; currently 1.1	§7
Boosting: depth / rate / rounds	4 / 0.05 / 400	§7
Count prior β_b (s / m / l / xl)	1.4 / 8.9 / 21.8 / 54.3	§5
Burden credibility $K_{B,b}$ (s / m / l / xl)	1.9 / 13.8 / 11.8† / 50.2†	§10
Grade cut-points (within band)	0.08 / 0.25 / 0.70 / 0.87 / 0.95	§10
Provisional grade cap	Satisfactory	§10
Validation holdout	carrier-disjoint, DOT mod 5 = 0	§13

† The large and xlarge K_B include reviewed scale factors (×0.25 and ×0.5 respectively): the §13 grade-monotonicity gate showed the unscaled constants over-shrink those bands, and the gate re-validates the setting on every refresh. All per-band constants above are re-estimated and logged on every run (`fred_v5_run_metrics.json`); the values shown are the June 2026 production refresh.

§B Severity-Weight Sensitivity

§4 frames the severity weights as a compressed harm index rather than a cost-proportional scale. A fair question follows: *how much do the published grades depend on the chosen fatal weight?* To answer it, the entire pipeline — burden targets, model fit, calibration, credibility, within-band percentiles — is re-run with the per-fatality weight moved from 12 to 8 and to 20, holding everything else fixed.

Table B.1 — Rank stability of grades under alternative fatal weights (June 2026 data).

Fatal weight	Spearman vs production	Worst band Spearman	Same grade	Within one grade
8 (vs 12)	0.996	0.995	96.9%	99.99%
20 (vs 12)	0.996	0.992	95.6%	99.96%

Carrier rankings are nearly invariant: rank correlations exceed 0.99 in every band, about 96% of carriers keep the identical letter grade, and well over 99.9% move at most one grade. The grade measures *which carriers* generate harm; the fatal weight mostly rescales *how much* harm — and the level is re-anchored by calibration regardless. Tail severity itself is priced by the dedicated fatal-crash model of §11.

This document specifies the production FRED scoring engine as of June 2026. Every formula, constant, and threshold is taken directly from the live scoring code; figures use either the stated production run or a faithful simulation of the named formula. The engine is re-fit and re-scored weekly on the full federal dataset.