A harmonized panel of United States compression-ignition engine certification data, with resolved certification lineage

Imafidon, Osariemen[1]

2026-09-15

[1] Independent Researcher. ORCID 0009-0006-3069-4674. Correspondence: odimafid@gmail.com.

Preprint. Not peer reviewed. Part of the FACET research program.

Abstract

The United States Environmental Protection Agency (EPA) publishes emission certification records for compression-ignition engines, but distributes them as two spreadsheets whose internal structures are mutually incompatible and whose true column headers are not on the first row. The result is a public record that is open in principle and costly to use in practice: every researcher must independently reverse-engineer two layouts, and the characteristic failure is not an error message but a quietly wrong number.

We present CIDEX, a harmonized engine-family-level panel reconciling both sources into a single schema: **8,627 engine families**, **9,794 engine configurations** and **94,869 emission records** from **108 manufacturers**, covering model years 2011–2027.

Beyond reorganisation, we resolve a structure present in the source but not previously published in usable form: **5,982 nonroad families declare a predecessor family**, and these declarations chain. Resolving them yields **2,541 distinct certification lineages** with a maximum depth of **14 model years**, allowing a family's regulatory history to be followed across more than a decade. We show the lineage graph is a forest, and that a naive traversal reports 30 spurious cycles arising from an undocumented self-reference convention.

We further identify 50 records carrying physically impossible negative certification results, and four model years whose apparent decline is an artefact of EPA's archive boundaries rather than a change in the fleet. Both are published as flagged data rather than silently corrected or removed.

CIDEX preserves certification units as issued and performs no silent conversion: highway results are in g/bhp-hr and nonroad results in g/kW-hr, and we argue that normalising them would convert a visible incompatibility into an invisible one.

Keywords: engine certification; compression-ignition engines; emissions; data harmonization; open government data; nonroad engines

---

1 Background and summary

1.1 The record and its unit

No compression-ignition engine may be sold in the United States without a certificate of conformity. The unit of certification is the **engine family** — a group of engines expected to share emission characteristics — rather than an individual engine model. A family is tested at one or more configurations, results are compared against the standard applicable to its service class and model year, and a certificate is issued.

The accumulated record is the only comprehensive public account of what engines were approved to emit, by whom, and under which standard. It bears directly on fleet composition, technology adoption, the realised stringency of successive standards, and the gap between certified and in-service behaviour.

1.2 Open but not usable

EPA publishes this record openly, without authentication, refreshed quarterly. The difficulty is structural, and it has three parts.

First, **both workbooks place their column headers on the second row**, reserving the first for a merged group header. The default behaviour of standard spreadsheet readers is to treat row one as the header, which yields a frame whose columns are named `Unnamed: 0` onward and whose first data row contains the real headers. No exception is raised.

Second, **the two files are organised in opposite directions**. The highway file is long on pollutant and wide on test type; the nonroad file is wide on pollutant beneath a two-level header, with family attributes distributed on *both* sides of the measurement block.

Third, **the two encode pollutants and units differently**, so any cross-cutting question requires a reconciliation that the publisher does not provide and no user has an obligation to perform correctly.

The consequence is that this work is duplicated by every user, documented by none, and its errors are invisible. Our contribution is to perform it once, publicly, with provenance and an explicit account of every judgement.

1.3 Contributions

1. A harmonized, analysis-ready panel covering both engine categories under one schema.
2. A resolved **certification lineage**, linking families across model years — present in the source as unresolved references, not previously published in usable form.
3. A characterisation of two defects in the published record: physically impossible values, and structural discontinuities at archive boundaries that masquerade as trends.
4. A reproducible pipeline with hash-level provenance and a verification procedure that blocks publication until an author has checked the output against the source by hand.

---

2 Methods

2.1 Sources

Two EPA workbooks, both US Government works without copyright restriction: heavy-duty highway gasoline and diesel engines (model years 2015–2026, 7,804 source rows) and nonroad compression-ignition engines (2011–2027, 10,214 rows).

EPA replaces these files in place each quarter with no changelog and no versioned archive. A dataset built from them is therefore only identified by the bytes it was built from, not by the URL. We record the SHA-256 digest of every source file; the deposited version carries fixed digests that a rebuilder can verify.

2.2 Schema

The target is a single long form keyed on **(panel, model year, engine family, pollutant, test type)**, with three supporting tables.

We establish that the engine family name is a **true key**: in both panels the count of distinct family names equals exactly the count of distinct (model year, family) pairs (702 and 7,925 respectively), because EPA encodes the model year in the name's first character. No surrogate key is introduced, so every row remains traceable to EPA's own record by a value a reader can look up.

We separate **family-level** from **configuration-level** attributes. In the nonroad source, 1,487 of 7,925 families occupy more than one row — up to 34 for a single family — differing by engine code, model, displacement or test procedure. Retaining these at family level would require an arbitrary row to win. The separation is enforced by assertion: the build fails if any attribute remaining in the family table varies within a family.

2.3 Harmonization

Highway. Read with the header on row two. The pollutant name maps through a controlled vocabulary of 10 terms; an unmapped value raises rather than being dropped, since a silently discarded pollutant is the precise failure this work exists to prevent. The test-type block is unpivoted, yielding one record per test type and removing 2,184 rows that carry no value in any measurement column as artefacts of the unpivot rather than facts about an engine.

Nonroad. The two-level header is resolved without positional assumptions, which would break on any column insertion. The group row is forward-filled, then a column is classified as a measurement **only if its sub-name is one its group actually contains**. This test is load-bearing: the seven attribute columns trailing the measurement block inherit a forward-filled group label and would otherwise be misclassified as pollutants. The classifier identifies 22 measurement and 24 attribute columns.

2.4 Lineage resolution

5,982 nonroad families name a predecessor. Resolution is a traversal to each family's root, with three cases handled explicitly.

Self-reference. 30 families name themselves. EPA appears to use this to denote an unchanged re-certification; it is undocumented. A naive traversal treats it as a cycle — our first implementation reported 109 — and we treat `parent(x) = x` as a terminus.

Orphans. 568 families name a predecessor absent from the published window, almost all pre-2011 families held in EPA's archive file. These lineages are truncated rather than erroneous, which means **a recorded depth of zero does not always denote the first of its line**. Both counts are published.

Cycles. Detected with a visited set. Count: 0. The graph is a forest.

2.5 Units

Highway results are certified in g/bhp-hr, nonroad in g/kW-hr, nonroad smoke opacity in percent. **No conversion is performed.** Each record carries its unit system explicitly.

We regard this as a substantive design decision rather than an omission. Normalising to a single system would produce a tidier table in which a user aggregating across panels obtains a plausible and wrong result with nothing to signal the error. Leaving the units divergent makes the incompatibility visible at the point of use. A conversion helper is provided and must be called deliberately; the margin helper raises rather than computing across unit systems.

---

3 Data records

Table	Rows	Contents
`cidex_family.csv`	8,627	One row per (panel, model year, family)
`cidex_config.csv`	9,794	Configurations within a family
`cidex_emissions.csv`	94,869	One row per family, pollutant, test type
`cidex_carryover.csv`	7,925	Resolved lineage: parent, root, depth

14 pollutants, 3 test types, 108 manufacturers. Parquet mirrors accompany each table. A machine-readable codebook documents every column.

`NMHCE` (Non-Methane Hydrocarbon Equivalent, 13 records) is retained as distinct from `NMHC` (14,219 records) because it is a different regulatory construct. Folding it in would destroy information a user cannot reconstruct; keeping it separate costs a filter expression.

---

4 Technical validation

4.1 Automated checks

10 checks execute on every build and all pass: key uniqueness, referential integrity between emission and family tables, non-null unit labelling, unit consistency within each panel, absence of cycles in the lineage graph, and complete flagging of anomalous values. These establish **internal consistency** and cannot establish correspondence with EPA's record.

4.2 Units established by evidence, not assertion

The highway source states its units nowhere; its headers are bare. That highway results are in g/bhp-hr was initially an inference from domain knowledge.

We replaced it with a test. US heavy-duty highway standards have known values in g/bhp-hr, and the modal certification standard per pollutant in the panel is compared against them:

Pollutant	Known standard (g/bhp-hr)	Modal value in CIDEX
NOx	0.2	0.2
PM	0.01	0.01
NMHC	0.14	0.14
CO	15.5	15.5

The same standards in g/kW-hr would be 0.268, 0.013, 0.188, 20.786, an order of magnitude distant in every case. The test cannot pass by coincidence and runs on every build.

4.3 Manual verification

No automated check can establish correspondence with the source. The pipeline ships five named spot-checks, each probing a distinct failure mode — the standard case, the archive boundary, the wide-to-long transformation, lineage resolution, and an anomalous value — which an author looks up in EPA's certificate tool and compares by hand.

Publication is blocked until this is signed. The attestation **pins which records were checked**, because the selector chooses a median row and its output moves when the row set changes; without pinning, a signature could come to refer to records nobody examined.

---

5 Anomalies in the published record

5.1 Physically impossible values

50 records (0.0527% of the panel, 18 families, all nonroad) carry a **negative** certification result. An emission rate cannot be negative.

We publish them unaltered, flagged. Correcting them would break the claim that CIDEX faithfully reorganises a published record, and would conceal a finding: either the source contains data-entry errors worth raising with EPA, or an encoding convention exists that neither party has documented. Both are more useful to a reader than a quietly cleaned column.

5.2 Archive boundaries that resemble trends

Both panels thin sharply at their edges, for two unrelated reasons. EPA holds older model years in **separate archive files**, so the earliest year of each panel contains only whichever families happen to sit in the current file. Separately, the newest model year is still being certified when the file is retrieved.

Panel	Model year	Families	Share of interior median
highway	2015	3	5%
highway	2016	29	45%
nonroad	2011	66	12%
nonroad	2027	8	2%

A time series spanning these years reads a filing artefact as a collapse or a boom. We flag every affected row rather than documenting the hazard in prose alone, and report usable ranges: highway MY2017–2026, nonroad MY2012–2026.

---

6 Usage notes

Three cautions govern correct use.

Do not compare across panels without converting units. Filter on `panel`, or convert deliberately.

Exclude incomplete years from time series. Filter `year_coverage == "complete"`.

Decide about flagged values explicitly. Filter on `quality_flag` if physically plausible values are required.

Beyond these, the lineage table supports questions the panel alone cannot answer: how long a certification family persists, how technology changes propagate through successive re-certifications, and how the distribution of lineage depth differs across power categories and tiers.

6.1 Limitations

EPA's archive files are not ingested in this version. Manufacturer names are not normalised across years. Nonroad records carry no pollutant-specific standard, because the source states it only at family level and inferring it would be a guess presented as data. Certification results are laboratory values on a certification fuel and duty cycle, and are not in-service emissions.

Data availability

The dataset is deposited on Zenodo; the DOI is recorded in the repository's CITATION.cff on release and is not yet assigned at the time of this preprint. The dataset is released under CC BY 4.0.

Code availability

The complete pipeline is at https://github.com/osariemenimafidon/cidex under the MIT licence, with a technical report, codebook, verification checklist and provenance log. A single command rebuilds every published figure from the source files and diffs the result against the deposited values.

Competing interests

The author declares no competing interests. This work received no funding.

References

1. US Environmental Protection Agency. *Annual Certification Data for Vehicles, Engines, and Equipment.* https://www.epa.gov/compliance-and-fuel-economy-data/annual-certification-data-vehicles-engines-and-equipment

2. US Environmental Protection Agency. *Control of Air Pollution from New Motor Vehicles and New Motor Vehicle Engines.* 40 CFR Part 86.
3. US Environmental Protection Agency. *Control of Emissions from New and In-Use Nonroad Compression-Ignition Engines.* 40 CFR Part 1039.
4. US Environmental Protection Agency. *Engine Testing Procedures.* 40 CFR Part 1065.

Preprint generated 2026-09-15. Every quantity is interpolated from the pipeline's own statistics file; none is transcribed by hand.