SmartFootPrintAI — Hybrid MRIO–LCA (CO_2, Land, Water): End-to-End Pipeline & QA

August 24, 2025

Abstract

This document specifies and audits the complete hybrid MRIO–LCA pipeline used in *SmartFootPrintAI* to compute sector-level environmental intensities for exactly three indicators: CO_2 (kg/), Land (m^2 year/), and Water (m^3/). It details inputs, unit conversions (with formulas), aggregations (conceptual and mathematical), access to EXIOBASE Q ($19 \times R \times S$), indicator selection, and the comparison methodology between the baseline (OLD) and micro-enhanced (NEW) sector intensities.

1 Concept and Goal

We link macroeconomic multi-regional input–output (MRIO; EXIOBASE 2022) with micro process life cycle assessment (LCA; e.g., Clark et al. 2022, WFLDB extracts) to obtain robust sector-level environmental intensities. We constrain indicators to exactly three (CO_2, Land, Water) to maintain unit consistency and reduce propagation of noise. Open Food Facts (OFF) products are translated, normalized to per-kg and per-euro, mapped to EXIO sectors, and aggregated to sectoral coefficients. Where high-quality micro values exist, they selectively override MRIO satellite intensities; otherwise we keep MRIO baselines. We then audit NEW vs. OLD.

Intuition. MRIO guarantees economy-wide system completeness (full upstream supply chains), while micro LCA provides process precision. Hybridization balances completeness and specificity.

2 Step-by-Step Pipeline (Files, Where, and Why)

Let I=19 indicators in EXIOBASE, R=189 regions, S=163 sectors. We propagate only the set $\mathcal{K} = \{CO_2, \text{Land}, \text{Water}\}$.

Detected Input/Project Files (example paths)

- `off_translated (3).parquet`, `product_to_sector_mapping.parquet`, EXIO zarr metadata, FAOSTAT CSVs, etc.

2.1 A. Data Pre-processing (Open Food Facts)

1. Translate product texts to English (cached Parquet; on-device model or prior cache).
2. Normalize text: lowercase, `unidecode`, strip, collapse spaces; deduplicate.
3. Output: `off_translated (3).parquet` (canonical product table).

2.2 B. OFF → CPC → EXIOBASE Mapping

1. Use precomputed mapping with confidence weights $w \in [0, 1]$; file: `product_to_sector_mapping.parquet`.
2. Keep top-1 EXIO sector per product using the highest confidence (also export low-confidence diagnostics).
3. Output: product→EXIO sector mapping with w.

2.3 C. Micro LCA & Price Normalization (Clark/WFLDB + FAOSTAT)

For each product p we expect, where available:

co2_per_kg$_p$ [kg/kg], land_per_kg$_p$ [m^2 year/kg], water_per_kg$_p$ [m^3/kg], eur_per_kg$_p$ [kg] from FAOSTAT / producer prices.

Convert prices: USD/tonne → USD/kg → EUR/kg:

$$\text{USD/kg} = \frac{\text{USD/tonne}}{1000}, \qquad \text{EUR/kg} = \text{USD/kg} \times (\text{USD} \to \text{EUR}).$$

Convert per-kg to per- for indicator $x \in \mathcal{K}$:

$$x\_\text{per\_eur}(p) \;=\; \frac{x\_\text{per\_kg}(p)}{\text{eur\_per\_kg}(p)}. \tag{1}$$

Outputs: `products_normalized_units.csv` (audited per-kg, EUR/kg, and per-).

2.4 D. Aggregate Product → Sector (Regionless NEW Q)

Let $\mathcal{P}_s$ be the set of products mapped to sector s, with confidence weights w_i. We compute weighted means for per- intensities:

$$\bar{x}_s \;=\; \frac{\sum_{i \in \mathcal{P}_s} w_i \, x\_\text{per\_eur}(i)}{\sum_{i \in \mathcal{P}_s} w_i}, \qquad x \in \mathcal{K}. \tag{2}$$

Outputs: `sector_micro_intensities.csv`, `Q_new_sector_regionless.csv` (CO_2/Land/Water per).

2.5 E. Access and Prepare OLD MRIO Q $(19 \times R \times S)$

We read EXIOBASE-2022 $Q \in \mathbb{R}^{I \times R \times S}$ and total outputs $T \in \mathbb{R}^{R \times S}$. To obtain sector-only, regionless per- values we use:

Method 1 (Output-weighted averaging).

$$w_{r|s} = \frac{T_{r,s}}{\sum_{r'} T_{r',s}}, \qquad q^{\text{global}}_{i,s} \;=\; \sum_{r=1}^{R} w_{r|s} \, Q_{i,r,s}. \tag{3}$$

Outputs: `Q_all19_global_regionless.csv` (19 indicators, per), and the three-indicator slice `Q_old_global_regionless.csv`.

2.6 F. Align, Impute, Compare OLD vs NEW

1. Reorder NEW sectors to EXIO order; preserve all S sectors.
2. Impute missing NEW per- values with the indicator mean across sectors; log flags.
3. Compute absolute/relative differences versus OLD.

Outputs: `Q_new_sector_aligned.csv`, `Q_new_sector_aligned_imputed.csv` (+.npy), `Q_compare_new_vs_ol` `Q_compare_new_vs_old_long.csv`, `Q_top5_diffs_by_indicator_UNSCALED.csv`.

3 Unit Conversions (Formulas)

Water volume. If data are in liters/kg, convert to m^3/kg:

$$\text{water\_per\_kg } [\text{m}^3/\text{kg}] \;=\; \frac{\text{water\_L\_per\_kg}}{1000}. \tag{4}$$

FAOSTAT price. Producer price conversion (year t):

$$\text{USD/kg}_t = \frac{\text{USD/tonne}_t}{1000}, \qquad \text{EUR/kg}_t = \text{USD/kg}_t \times (\text{USD} \to \text{EUR})_t. \tag{5}$$

Per- intensities. For $x \in \mathcal{K}$:

$$x\text{\_per\_eur} = \frac{x\text{\_per\_kg}}{\text{eur\_per\_kg}}, \tag{6}$$

with units: CO_2 [kg/], Land [m^2 year/], Water [m^3/].

MRIO per-. Using either output weights or divide-by-T yields sector-only MRIO intensities.

4 Aggregations: Concepts and Equations

4.1 Product → Sector

Weighted means with mapping confidence w_i.

4.2 Region → Global Sector

Use either output weights or divide-by-T; under consistent currency, both are equivalent.

4.3 Imputation

For each $x \in \mathcal{K}$, fill NaNs in $\bar{x}_s$ with the mean across sectors; record flags.

5 Resulting Outputs (File Catalog)

- `products_normalized_units.csv` — product-level, audited units (per-kg, EUR/kg, per-).
- `sector_micro_intensities.csv` — diagnostics: per-kg, per-, counts per sector.
- `Q_new_sector_regionless.csv` — NEW sector-only Q (CO_2/Land/Water per).
- `Q_new_sector_aligned.csv` — NEW, aligned to EXIO order.
- `Q_new_sector_aligned_imputed.csv` (+ .npy) — NEW with mean imputation + flags.
- `Q_all19_global_regionless.csv` — OLD, 19 indicators aggregated to sector (per).
- `Q_old_global_regionless.csv` — OLD, selected CO_2/Land/Water.
- `Q_old_divT_global_regionless.csv` — OLD via divide-by-T (per).
- `Q_compare_new_vs_old.csv` — wide comparison (old/new/absΔ/relΔ).
- `Q_compare_new_vs_old_long.csv` — tidy long version.
- `Q_top5_diffs_by_indicator_UNSCALED.csv` — top-5 absolute diffs per indicator.

6 Accessing EXIOBASE Q and Indicator Choice

We access EXIOBASE-2022 $Q \in \mathbb{R}^{I \times R \times S}$ with $(I, R, S) = (19, 189, 163)$ and $T \in \mathbb{R}^{R \times S}$. Indicators are chosen by fixed indices: $CO_2 \to 7$, Land $\to 3$, Water $\to 2$. We aggregate to sector-only, regionless per- via output-weighted averaging or divide-by-T.

7 Matrix Comparison Methodology

Let $Q^{\mathrm{new}} \in \mathbb{R}^{3 \times S}$ and $Q^{\mathrm{old}} \in \mathbb{R}^{3 \times S}$ be aligned by sector. For indicator $k \in \{1, 2, 3\}$ (CO_2, Land, Water) and sector s:

$$\Delta_s^{(k)} = Q_{k,s}^{\mathrm{new}} - Q_{k,s}^{\mathrm{old}}, \tag{7}$$

$$\delta_s^{(k)} = \frac{\Delta_s^{(k)}}{Q_{k,s}^{\mathrm{old}}} \quad \text{(guard division-by-zero).} \tag{8}$$

We report coverage, summary statistics (mean, median, p90, max), and the top-5 sectors by $|\Delta|$ per indicator.

8 QA and Diagnostics

- Unit sanity checks: liters$\to$m^3; price construction; per- recomputation.
- NaN scans and imputation flags for NEW per-.
- Coverage: products with LCA; products with prices; sector coverage after aggregation.
- MRIO shapes/currency: confirm Q per (or M) and rescale as needed.
- Consistency: sector order alignment; numeric types; no silent coercions.