Construction and Validation of the Monthly Mix-Adjusted Labor Productivity Panel (`LP_panel`)

Project: UC Berkeley Econ H191 — Israel Ports Reform
Data product: `LP_panel.tsv` (unified long panel: port+terminal, monthly)
Build time: 2025-10-14 (UTC)
Code: `Data/LP/build_lp_mixadjusted.py` (hash recorded in meta)
Inputs: TEU by port (monthly slice), mixed Tons (All-Ports/port/terminal), terminal-year KPIs (Π) and L proxy (terminal×month)
Outputs: Port-month LP (analysis), terminal-month LP (extension), identity LP (diagnostic), unified panel, QA table, META JSON

Executive Abstract

We construct a **monthly** labor-productivity (LP) series at the **port** level (primary) and **terminal** level (extension) from heterogeneous public sources. The series preserves **annual productivity levels** observed in official KPIs (Π) but gains **within-year monthly variation** through a **cargo-mix/weight factor** derived from **tons per TEU**. The build follows strict, documented precedence rules to form **port-month tons**; it **inherits monthly TEU exactly** (never disaggregates quarterlies); and it avoids the classic **mechanical endogeneity** that would arise if we used TEU/L as the outcome. The process includes QA that enforces **key uniqueness**, verifies **annual preservation**, and reports the **coverage mix** of tons sources. The unified output `LP_panel.tsv` is reproducible (CLI + hashes), audit-ready, and directly usable for the Econometrics plan ($Y_{pt} = \ln LP_{pt}$ with port and time FE).

Part I — Conceptual Design & Econometric Validity

I.1 Objectives & Constraints

Objective. Measure monthly labor productivity for Israel's container ports to evaluate the effects of competition (Bayport, HCT) and privatization.
Grains. (a) **Port×month** (primary outcome for econometrics); (b) **Terminal×month** (extension); (c) unified long panel.
Constraints.

- TEU source mixes **monthly** and **quarterly** entries by port. We **only use monthly TEU**; we do **not** fabricate monthly TEU by splitting quarterlies.
- Tons data are **mixed-grain** (All-Ports totals, port totals, terminal rows).
- KPIs (Π = TEU per work-hour) are **annual by terminal**; **no monthly Π** exist.
- Labor proxy `L` (terminal×month) exists and is **calibrated to Π annually** but should **not** mechanically determine our outcome.

I.2 General Approach (Intuition)

We decompose “monthly productivity” into two orthogonal components:

1. **Annual levels from Π (terminal-year KPIs).** At the **port**, we aggregate terminal annual Π using **quarter-constant terminal shares** of throughput to form a **port-year mix-base Π** (no within-year smoothing of Π).
2. **Monthly variation from work content (tons/TEU).** We compute a **port-month** ratio `tons/TEU`, winsorize outliers **within port-year**, and **re-base** it so the **annual mean equals 1**. This yields a **monthly mix factor** w that scales the annual Π level up or down each month.

LP definition. The **analysis outcome** is $\mathrm{LP}_{\mathrm{mix}} = \Pi_{\mathrm{mixbase}} \times w$. Annual means equal Π_{mixbase} by construction, while **within-year** movements reflect contemporaneous cargo mix/weight, not labor measurement artifacts.

I.3 Why This Approach

- **Granularity without invention.** Using observed **monthly tons and TEU** lets us induce legitimate monthly variation through w ; we never synthesize monthly TEU from quarterly figures.
- **Measurement integrity.** Annual calibration comes from official KPIs (Π). Monthly dynamics come from w . We do **not** use `TEU/L` (identity) as the analysis series.
- **Comparability across ports and time.** Re-basing w to mean 1 by port-year preserves levels and removes composition differences that are constant within the year.
- **Transparency and auditability.** A deterministic precedence for tons, canonical name mapping, and a coverage report (which source supplied each month) make assumptions explicit and testable.

I.4 Bias Control & Econometric Validity

- **Avoids mechanical endogeneity.** Using $\mathrm{LP}_{\mathrm{mix}} = \Pi \times w$ (instead of TEU/L) decouples the outcome from the constructed L proxy. L affects the identity diagnostic only.
- **Annual preservation.** Because $E[w \mid p, y] = 1$, the annual mean of $\mathrm{LP}_{\mathrm{mix}}$ equals the Π mix-base. No spurious level shifts are introduced.
- **Outliers contained, not distorted.** Winsorizing `tons/TEU` at 1–99% by port-year stabilizes w but, due to the re-base, **does not** alter annual means.
- **All-Ports allocation symmetry.** When only system-wide tons exist, we allocate to **Ashdod, Haifa, Eilat** proportionally to **that month’s TEU shares**. This is symmetric and used **only** when port totals and terminal sums are unavailable. The QA table quantifies how often allocation was required.
- **Backfill transparency (Π for new terminals).** Any upstream Π backfills (for early Bayport/HCT years) affect **annual levels** only; **within-year** variation still comes from w . Fixed effects and event-time designs absorb level differences that are not aligned with treatment timing.

Part I-b — Formal Mathematics

Let ports $p \in \{\text{Ashdod}, \text{Haifa}, \text{Eilat}\}$; terminals $i \in p$; month m ; year $y = y(m)$; quarter $q = q(m)$.
Inputs: $TEU_{p,m}$ (monthly), mixed tons $tons_{All,m},\ tons_{p,m}^{(port)},\ tons_{i,m}^{(term)}$, annual terminal $\Pi_{i,y}$.

(1) Port-month Tons via Precedence

1. If a **port total** exists: $tons_{p,m} = tons_{p,m}^{(port)}$.
2. Else, if **terminal rows** exist: $tons_{p,m} = \sum_{i \in p} tons_{i,m}^{(term)}$.
3. Else, if **All-Ports** total exists: allocate across $\mathcal{P} = \{\text{Ashdod}, \text{Haifa}, \text{Eilat}\}$ by same-month TEU shares:

$$ tons_{p,m} = tons_{All,m}\cdot \frac{TEU_{p,m}}{\sum_{p'\in \mathcal{P}} TEU_{p',m}}. $$

4. Else: $tons_{p,m} = \text{NA}$.

Record `tons_source` $\in \{\text{port\_total}, \text{sum\_terminals}, \text{allocated\_allports}, \text{no\_source}\}$.

(2) Monthly Mix Factor from Tons/TEU

Raw ratio $r_{p,m} = \begin{cases} \dfrac{tons_{p,m}}{TEU_{p,m}}, & TEU_{p,m} > 0 \\ \text{NA}, & \text{otherwise} \end{cases}$.

Winsorize within (p, y) at $[1\%, 99\%]$: $r_{p,m}^{W}$. \
Re-base to mean 1 by (p, y) :

$$ w_{p,m} = \frac{r^W_{p,m}}{\dfrac{1}{M_{p,y}}\sum_{m\in y} r^W_{p,m}},\quad \text{with fallback } w_{p,m}=1 \text{ if denominator }=0\text{ or NA}. $$

Property: $\dfrac{1}{M_{p,y}} \sum_{m \in y} w_{p,m} = 1.$

(3) Port-Year Π Mix-Base from Terminal Π

Quarter-constant terminal shares $s_{i,p,q}$ from terminal TEU:

$$ s_{i,p,q} = \frac{\sum_{m\in q} TEU_{i,m}}{\sum_{j\in p}\sum_{m\in q} TEU_{j,m}}, \quad (s_{i,p,q}=0 \text{ in pre-opening months}). $$

Monthly port mix-base:

$$ \Pi^{\text{mix}}*{p,y(m)} = \sum*. $$} s_{i,p,q(m)}\, \Pi_{i,y(m)}

(4) Monthly LP Series

Port (analysis): $LP^{\text{mix}}_{p,m} = w_{p,m} \cdot \Pi^{\text{mix}}_{p,y(m)}$. \
Port (diagnostic identity): $LP^{\text{id}}_{p,m} = \dfrac{TEU_{p,m}}{L_{p,m}}$. \
Terminal (extension): $LP^{\text{mix}}_{i,m} = w_{p(i),m} \cdot \Pi_{i,y(m)}$ (NA in pre-opening/non-operating months).

(5) Annual Preservation (Sketch)

$$ \frac{1}{M_{p,y}} \sum_{m\in y} LP^{\text{mix}}*{p,m} = \left(\frac{1}{M}\right)\sum_{m\in y} w_{p,m}\right) \Pi^{\text{mix}}*{p,y} = 1\cdot \Pi^{\text{mix}}*{p,y}. $$

Part II — Implementation & Provenance (Code-Level)

II.1 Inputs (Paths, Schemas, Keys)

1) **TEU (port×time; monthly+quarterly mixed)**
- Required columns: `Port`, `Year`, `Month` *or* `Period` / `MonthIndex`; either `TEU` or `TEU_thousands`; optional `Freq` with values like `"Monthly"` or `"M"`.
- Processing: filter to **monthly** rows (tolerant to labeling), derive `(year, month)` if needed, convert `TEU_thousands×1000 → TEU`, canonicalize ports {Ashdod, Haifa, Eilat, All Ports}, group to unique `(port,year,month)`.

2) **Tons (mixed: All-Ports / port / terminal)**
- Required columns: `PortOrTerminal`, `Month-Year`, `tons_k`.
- Processing: parse `Month-Year → (year, month)`, canonicalize **ports** (incl. `All Ports`) and **terminals** (`Haifa SIPG/Bayport → Haifa-Bayport`, `Ashdod HCT/Southport → Ashdod-HCT`, `* Legacy`), convert `tons_k×1000 → tons`.

3) **L proxy / KPIs (terminal×month)**
- Required columns at minimum: `port, terminal, year, month, quarter, TEU_i_m, L_hours_i_m, Pi_teu_per_hour_i_y`.
- Processing: canonicalize names, derive `month_index`, `operating` flag (TEU_i_m>0 & L_hours_i_m>0).

II.2 Transform Graph (Provenance Map)

A. TEU base (monthly) → `(port,year,month, TEU_p_m)`
B. Tons precedence → `(port,year,month, tons_p_m, tons_source, compare_diff)`
C. Mix factor → `(port,year,month, tons_per_teu, w_p_m)`
D. Π mix-base → `(port,year,month, Pi_p_y_mixbase)`
E. Port LP → `(port,year,month, LP_port_month_mix)` + identity (diagnostic) + `l_port_m`

F. Terminal LP → `(port,terminal,year,month, LP_term_month_mixadjusted)`
G. Unified panel → `LP_panel.tsv` (stack E + F with a common schema)
H. QA → `qa_lp_report.tsv` (uniqueness, annual preservation, w dispersion, tons_source coverage)
I. META → `_meta_lp_mixadjusted.json` (paths, sha256, params, row counts)

II.3 Detailed Steps (What the Code Does)

Step 1 — Load & Normalize

- **TEU loader** tolerates TSV/CSV, harmonizes headers, interprets `Freq` (`Monthly` or `M`), or infers monthly by presence of `Month` / `MonthIndex` /parsable `Period`.
- **Tons loader** parses dates, maps `PortOrTerminal` to `port_label` and `terminal_label`, and scales `tons_k`.
- **L loader** standardizes column names, canonicalizes terminals and ports, and sets `operating`.

Step 2 — Build Port-Month Tons (Precedence Engine)

From the TEU base (restrict to {Ashdod, Haifa, Eilat}): 1. Merge **port totals** by `(port,year,month)`.
2. Build **terminal sums** by mapping canonical terminals to their **parent port** and summing tons.
3. Merge **All-Ports totals** by `(year,month)`. 4. Compute **allocation denominator**: `teu_alloc_sum = Σ TEU_p_m` over the allocation set `{Ashdod, Haifa, Eilat}` for each month. 5. Apply precedence: choose `tons_porttotal` if present; else `tons_terminalsum`; else allocate `All-Ports` by `TEU_p_m/teu_alloc_sum` for each port; else `tons_p_m=NA`.
6. Record `tons_source` and `compare_diff` when both sources exist.

Step 3 — Compute the Mix Factor *w*

- Ratio: `tons_per_teu = tons_p_m / TEU_p_m` where `TEU_p_m>0`, else NA.
- Winsorize **within (port,year)** at `[1%, 99%]` on finite values.
- Re-base to mean 1 within each (port,year); if a group mean is `0`/NA (degenerate), set `w_p_m=1`.
- Output carries `tons_source` forward for diagnostics.

Step 4 — Π Mix-Base at the Month Grain

- Aggregate `TEU_i_m` to terminal-**quarter** totals; compute **quarter shares** `s_{i,p,q}` within each port.
- Multiply shares by **terminal annual Π** and sum across terminals to obtain `Pi_p_y_mixbase` at the month grain.

Step 5 — LP Series

- **Port LP (analysis):** `LP_port_month_mix = w_p_m × Pi_p_y_mixbase`.
- **Identity diagnostic:** build `l_port_m = Σ L_hours_i_m` and `LP_port_month_id = TEU_p_m / l_port_m` (for QA/triangulation only).
- **Terminal LP (extension):** `LP_term_month_mixadjusted = w_{p(i),m} × Pi_i_y`; force **NA** in non-operating months to avoid structural zeros.

Step 6 — Unified Long Panel

Stack port rows and terminal rows into the **same column layout**: - Common fields: `level, port, terminal, year, month, month_index, quarter, TEU, tons, tons_per_teu, w, Pi, L_hours, LP_mix, LP_id, tons_source`. - For terminals, `tons*` and `LP_id` are NA by design; for ports, all fields are populated.

Step 7 — QA & META

- **Uniqueness checks** on keys (port×month; terminal×month).
- **Annual preservation:** report `mean(LP_mix)` vs `Pi_mixbase` by port-year and relative error.
- **w dispersion:** CV of `w` by port-year.
- **Coverage by tons_source:** counts of `port_total`, `sum_terminals`, `allocated_allports`, `no_source` for each port-year (flag fail if any `no_source>0`).
- **META JSON:** script path & timestamp, input paths with **sha256**, parameters (`winsor_pct`, `allocation_ports`), and **row counts** for all artifacts.

II.4 Invariants & Contracts (What Must Hold)

- **No duplicate keys.** `(port,year,month)` unique for port outputs; `(port,terminal,year,month)` unique for terminal outputs and for the unified panel.
- **Annual preservation.** For every (port,year), mean over months of `LP_port_month_mix` equals `Pi_mixbase` (within float tolerance).
- **TEU inheritance.** Monthly port TEU are **not** synthesized; they are inherited from the TEU source exactly.
- **NA policy.** If tons cannot be formed (no source) or `TEU_p_m≤0`, then `tons_per_teu` is NA; `w` defaults to 1 only when the **group mean** is undefined, not to conceal missingness.

II.5 Edge Cases & Safeguards

- **Zero/near-zero TEU months.** Ratios set to NA to prevent inf/NaN; group re-base yields `w=1` only if the group mean is undefined.
- **Terminal name variants.** Robust mapping (e.g., "Haifa SIPG", "Bayport" → `Haifa-Bayport`; "Ashdod HCT", "Southport" → `Ashdod-HCT`; `* Legacy`).
- **All-Ports allocation set.** Default includes **Ashdod, Haifa, Eilat**; configurable for sensitivity.
- **Port-vs-terminal disagreement.** If both exist and differ, keep **port totals**; log relative difference for review.

II.6 Reproducibility (CLI, Hashes, Row Counts)

Command (repo root):

```
python Data/LP/build_lp_mixadjusted.py
  --teu Data/Output/teu_monthly_plus_quarterly_by_port.tsv
  --tons Data/Output/monthly_output_by_1000_tons_ports_and_terminals.tsv
```

```
  --l-proxy Data/L_proxy/L_Proxy.tsv
  --allocation-ports Ashdod Haifa Eilat
```

Parameters: winsorization = `[0.01, 0.99]` by **port-year**; allocation ports = `{Ashdod, Haifa, Eilat}`.
Meta records: script path & time, **sha256** of each input, and the following counts from the run: - TEU rows: **664**; Tons (mixed) rows: **899**; L_proxy rows: **264**
- Port-month tons (`tons_pm`) rows: **498**; w rows: **498**
- Port LP rows: **498**; Terminal LP rows: **264**
- Unified panel rows: **762**; QA rows: **131**

(If any of these differ in a re-run, the META will reflect the new numbers; always archive META alongside outputs.)

II.7 Outputs (Schemas)

A) `LP_port_month_mixadjusted.tsv`

`port, year, month, month_index, TEU (teu_p_m), tons (tons_p_m), tons_per_teu, w (w_p_m), Pi_mixbase (pi_p_y_mixbase), LP_port_month_mix, tons_source, l_port_m, LP_port_month_id`

B) `LP_terminal_month_mixadjusted.tsv`

`port, terminal, year, month, month_index, quarter, operating, Pi_i_y (pi_teu_per_hour_i_y), w (w_p_m), TEU_i_m, L_hours_i_m, LP_term_month_mixadjusted`

C) `LP_port_month_identity.tsv` (diagnostic)

`port, year, month, TEU, l_port_m, LP_port_month_id`

D) `LP_panel.tsv` (unified long panel)

`level ∈ {port, terminal}, port, terminal, year, month, month_index, quarter, TEU, tons, tons_per_teu, w, Pi, L_hours, LP_mix, LP_id, tons_source`

E) `qa_lp_report.tsv`

Rows for: key uniqueness, annual preservation by port-year (with relative error), `w` CV by port-year, coverage by `tons_source` per port-year.

F) `_meta_lp_mixadjusted.json`

Script path/time, input paths + **sha256**, parameters, row counts for all artifacts.

II.8 Validation & Readiness Checklist

1) **Uniqueness:** zero duplicates on keys in all outputs.
2) **Annual preservation:** relative errors ≈ 0 in QA.
3) **Coverage:** zero `no_source` months; reasonable shares of `port_total`/`sum_terminals` vs `allocated_allports`.
4) **Non-flatness:** `w` CV by port-year is non-trivial (not degenerate).
5) **Span:** months written cover the intended windows around competition go-live and privatization.

II.9 Limitations & Robustness Menu

- **Dependence on monthly TEU coverage.** The panel's horizon is limited by the monthly slice in the TEU file; we do not synthesize from quarterly totals.
- **All-Ports allocation.** Symmetric, auditable, and used as a last resort. Robustness: drop allocated months or re-run with alternative allocation sets.
- **Winsorization level.** 1–99% is conservative; test 2–98% as sensitivity.
- **Π backfill (if present upstream).** Affects levels only; use FE and event-time; report sensitivity to α-bands if needed.

Part III — Ready-to-Use Notes for Econometrics

- **Outcome variable:** For Part B/C, use `Y_pt = ln(LP_mix)` from `level="port"` rows in `LP_panel.tsv`.
- **Fixed effects:** Include port FE and time FE per design; optionally add shock dummies (COVID, 2023Q4–2024).
- **Spillovers:** Construct spillover indicators as per the econometrics plan.
- **Diagnostics:** Plot identity `LP_id` vs `LP_mix` at low frequency; inspect QA `tons_source` shares by port-year.
- **Reproducibility:** Archive outputs with `_meta_lp_mixadjusted.json` and `qa_lp_report.tsv`.

Appendix A — Glossary

- **Π (Pi):** TEU per work-hour (annual, by terminal).
- **L:** labor hours proxy (terminal×month).
- **LP_mix:** monthly labor productivity (analysis series), `Π_mixbase × w`.
- **w:** monthly mix factor from winsorized `tons/TEU`, rebased to mean 1 by port-year.
- **tons_source:** provenance of port-month tons (`port_total`, `sum_terminals`, `allocated_allports`, `no_source`).

Appendix B — Quick Sanity Plots (suggested)

1) `w_p_m` by month within each port-year (line), to show non-flatness.
2) Annual mean of `LP_port_month_mix` vs `Pi_mixbase` (scatter 45°) by port-year (preservation).
3) Stacked bars of `tons_source` shares by port-year.
4) `LP_mix` vs `LP_id` (12-m moving average) for triangulation.

End of Report.