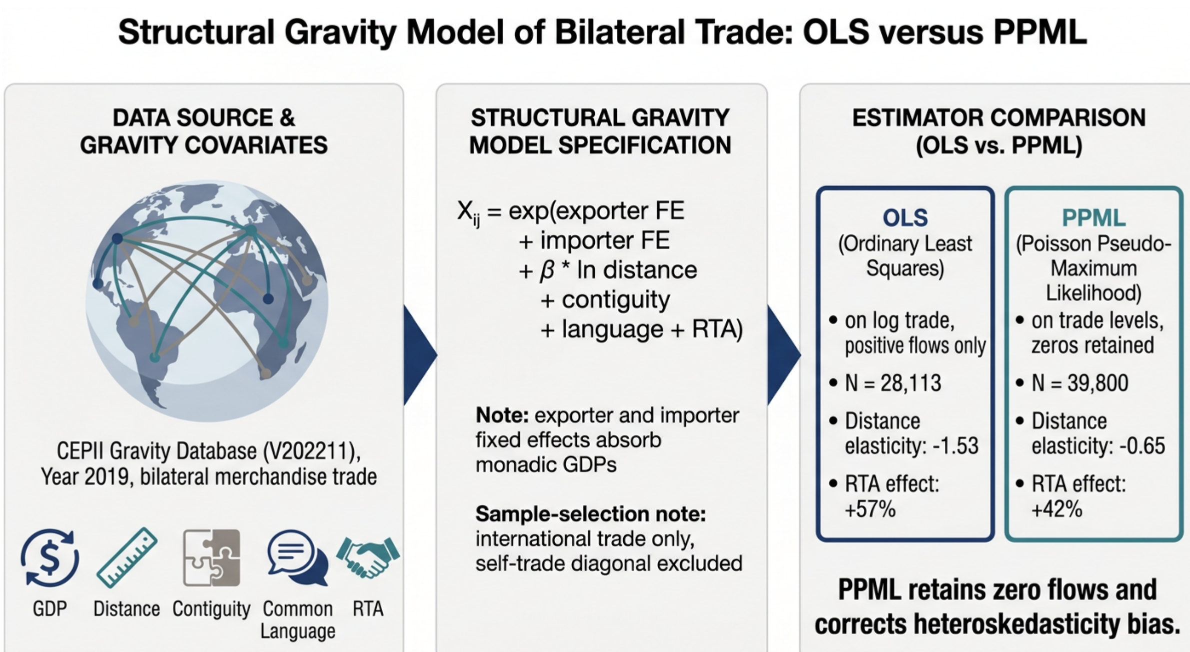
# Structural Gravity Model of Bilateral Merchandise Trade

## OLS versus Poisson Pseudo-Maximum-Likelihood, 2019

*Technical Report*

A structural gravity estimation using the CEPII Gravity Database (V202211)





**Author:** K-Dense Web
**Contact:** contact@k-dense.ai

**Date:** July 10, 2026
**Dataset:** CEPII Gravity Database, release V202211 (year 2019)

Fully reproducible: data acquisition and estimation code, regression tables (CSV), and figures accompany this report.

# Contents

# 1 Executive Summary

This report estimates a structural gravity model of bilateral merchandise trade for the single cross-section of the year 2019 and documents how the choice of estimator—ordinary least squares (OLS) on log trade versus Poisson pseudo-maximum-likelihood (PPML) on the level of trade—materially changes the headline elasticities. The gravity model is the empirical workhorse of international economics: it relates bilateral trade to the economic size of the two partners and to the bilateral frictions between them (distance, a shared border, a common official language, and membership in a regional trade agreement). The data are drawn from the **CEPII Gravity Database, release V202211** (Conte et al., 2022), restricted to 2019, and the specification includes exporter and importer fixed effects that control for each country's multilateral resistance and, by construction, absorb the monadic (country-specific) gross domestic products (GDPs).

We estimate the same specification two ways. The first is OLS on the natural logarithm of trade, which necessarily discards the roughly one-third of country pairs that record zero trade. The second is PPML on the level of trade, which retains those zero flows and is robust to the heteroskedasticity that pervades trade data (Santos Silva and Tenreyro, 2006). A critical sample-construction decision—the exclusion of self-trade (the diagonal where the exporter equals the importer)—was required to obtain a credible PPML distance elasticity; retaining those observations biased the elasticity toward zero.

**Headline results (year 2019, CEPII V202211)**

**OLS** (log trade, positive flows only): $N = 28{,}113$; distance elasticity $= -1.528$; RTA effect $= +57.2\%$; $R^2 = 0.764$.

**PPML** (trade level, zeros retained): $N = 39{,}800$ (28,113 positive + 11,687 zeros); distance elasticity $= -0.648$; RTA effect $= +42.0\%$; pseudo-$R^2 = 0.878$.

The estimator matters: OLS overstates the distance elasticity by a factor of about 2.4 and the RTA effect by roughly a third, because the log transformation, combined with heteroskedasticity and the deletion of zeros, biases the log-linear model (Jensen's inequality).

The empirical takeaway is unambiguous and consistent with the modern trade-econometrics consensus: PPML is the preferred estimator because it treats zero trade flows consistently and delivers coefficients that are robust to heteroskedasticity, whereas log-OLS is contaminated by a bias whose sign and size depend on the (unknown) pattern of the error variance. The methodological takeaway is equally important: careful sample construction—in particular, the exclusion of unobserved internal trade—is a precondition for interpreting the distance elasticity structurally.

# 2 Introduction

## 2.1 The gravity model of trade

The gravity model is the most successful empirical relationship in international economics. In its original, atheoretical form, introduced by Tinbergen (1962), bilateral trade between two countries is proportional to the product of their economic masses (GDPs) and inversely proportional to the distance between them, in direct analogy to Newton's law of gravitation. For four decades the gravity equation was estimated as a log-linear regression and criticized for lacking theoretical foundations. That criticism was decisively answered by Anderson and van Wincoop (2003), who derived the gravity equation from a general-equilibrium model with constant-elasticity-of-substitution (CES) preferences and showed that bilateral trade depends not only on bilateral trade costs but also on *multilateral resistance*—the average trade barriers each country faces with all of its partners. This structural interpretation is what turns a descriptive regression into a tool for counterfactual policy analysis.

The structural gravity equation for exports $X_{ij}$ from exporter $i$ to importer $j$ can be written as

$$X_{ij} \;=\; \frac{Y_i\,E_j}{Y}\left(\frac{t_{ij}}{\Pi_i P_j}\right)^{1-\sigma}, \tag{1}$$

where $Y_i$ is the value of production in $i$, $E_j$ is expenditure in $j$, $Y$ is world output, $t_{ij}$ is the bilateral trade cost, $\sigma > 1$ is the elasticity of substitution, and $\Pi_i$ and $P_j$ are the outward and inward multilateral-resistance terms. The central methodological insight of the modern literature—formalized by Fally (2015) and popularized by Head and Mayer (2014) and Yotov et al. (2016)—is that in a cross-section the multilateral-resistance terms, together with the size variables $Y_i$ and $E_j$, are fully captured by a set of *exporter and importer fixed effects*. This is the specification we adopt.

### 2.2 Estimator choice: the log-of-gravity problem

Because the theoretical equation (1) is multiplicative, the historically dominant practice was to take logarithms and estimate a linear model by OLS. Santos Silva and Tenreyro (2006), in their influential paper "The Log of Gravity," showed that this practice is generally inconsistent. Two problems arise. First, the log transformation is undefined for the many country pairs that trade nothing, forcing the researcher to drop zeros—a non-random selection that discards economically meaningful information. Second, and more subtly, the expectation of the logarithm of a random variable is not the logarithm of its expectation (Jensen's inequality), so when the error term is heteroskedastic—as trade data overwhelmingly are—the log-linear coefficients are biased even on the positive flows. Their proposed solution, PPML, estimates the multiplicative model directly on the level of trade, retains zeros naturally, and is consistent under weak assumptions. In the fifteen years since, PPML has become the default estimator for structural gravity (Santos Silva and Tenreyro, 2022; Head and Mayer, 2014; Yotov et al., 2016), with efficient high-dimensional-fixed-effects implementations now standard (Correia et al., 2020).

### 2.3 Scope of this report

This report locates a public gravity dataset that supplies, for country pairs by year, bilateral merchandise trade together with the standard gravity covariates, restricts it to 2019, and estimates the structural gravity model of bilateral exports both ways. We report, for each estimator, the distance elasticity, the RTA coefficient expressed as a percentage trade effect, the number of observations, and the (pseudo-)$R^2$; we then discuss the difference the estimator makes and why. The dataset is the **CEPII Gravity Database, release V202211** (Conte et al., 2022), whose bilateral trade flows are built from the CEPII BACI database (Gaulier and Zignago, 2010).

## 3 Data

### 3.1 Dataset name and version

**Dataset**
**CEPII Gravity Database**, release **V202211** (2022 release), restricted to the year **2019**. Source: `https://www.cepii.fr/DATA_DOWNLOAD/gravity/data/Gravity_csv_V202211.zip`. Documentation: Conte et al. (2022).

The CEPII Gravity Database is a “squared” panel in which every observation is an exporter–importer–year triplet, covering all countries over 1948–2020. It bundles bilateral trade flows with the full set of gravity covariates—geographic distance, contiguity, common language, colonial history, and regional-trade-agreement membership—together with macroeconomic aggregates such as GDP. The full release contains 4,699,296 rows across all years; streaming and filtering to 2019 yields 63,504 country-pair rows (a $252^2$ grid across 243 active reporters).

The primary dependent variable is `tradeflow_baci`, bilateral merchandise trade in thousands of current US dollars, sourced from the CEPII BACI reconciliation of UN COMTRADE mirror flows (Gaulier and Zignago, 2010). Table 1 lists the variables used.

**Table 1:** Variables retained from the CEPII Gravity Database (V202211), year 2019.

| **Column** | **Definition** |
|---|---|
| `iso3_o`, `iso3_d` | Exporter / importer ISO3 country codes |
| `tradeflow_baci` | Bilateral merchandise trade, BACI (thousand current USD) — dependent variable |
| `gdp_o`, `gdp_d` | Exporter / importer GDP (current USD) |
| `dist` | Bilateral distance (km, population-weighted) |
| `contig` | Contiguity dummy (= 1 if the pair shares a land border) |
| `comlang_off` | Common official language dummy |
| `rta` | Regional trade agreement membership dummy (from `fta_wto`) |

**Naming note.** Release V202211 does not contain a plain `rta` column; the standard RTA/FTA membership indicator in this release is `fta_wto` (equal to one if the pair is covered by a free-trade agreement notified to the WTO). We read `fta_wto` and rename it to `rta` for downstream consistency.

### 3.2 Treatment of zero and missing trade flows

A defining feature of the BACI trade data is that no-trade country pairs are encoded as *missing*, not as literal zeros. In the 2019 cross-section there are 31,120 positive flows (49.0%), zero literal zeros, and 32,384 missing flows (51.0%). Of the missing flows, 13,456 are *candidate structural zeros*: pairs whose trade is unobserved but whose core covariates (both GDPs and distance) are present. For the PPML step these candidate structural zeros—where covariates are complete—are treated as zeros, so that the extensive margin of trade (whether a pair trades at all) is retained in the estimation rather than silently discarded.

## 4 Methodology

Figure 1 summarizes the full estimation pipeline, from the raw multi-year database to the two regression tables.

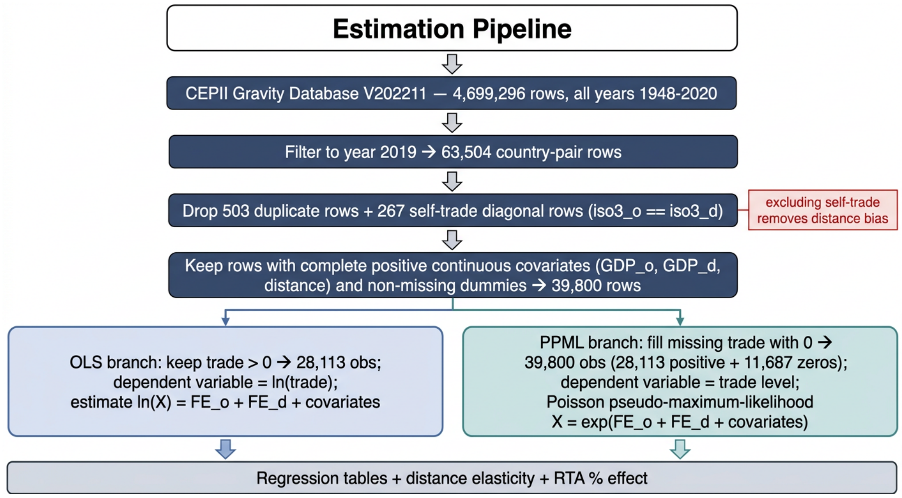



**Figure 1:** Estimation pipeline. The 2019 cross-section is cleaned of duplicate and self-trade rows and restricted to covariate-complete observations, then estimated two ways: OLS on log trade (positive flows only) and PPML on the level of trade (structural zeros retained).

## 4.1 Econometric specification

Let $i$ index the exporter and $j$ the importer. The structural gravity model of bilateral exports $X_{ij}$ is

$$\mathbb{E}[X_{ij}] \;=\; \exp\Big(\mu_i + \nu_j + \beta_{\text{dist}} \ln \text{dist}_{ij} + \beta_{\text{contig}}\, \text{contig}_{ij} + \beta_{\text{lang}}\, \text{comlang}_{ij} + \beta_{\text{rta}}\, \text{rta}_{ij}\Big), \tag{2}$$

where $\mu_i$ is an exporter fixed effect, $\nu_j$ is an importer fixed effect, and the bilateral covariates are log distance, contiguity, common official language, and RTA membership. The two estimators differ in how equation (2) is taken to the data.

### 4.1.1 OLS on log trade

Taking logarithms of the multiplicative model and adding an error term gives the log-linear specification, estimated on the sub-sample of *positive* flows:

$$\ln X_{ij} \;=\; \mu_i + \nu_j + \beta_{\text{dist}} \ln \text{dist}_{ij} + \beta_{\text{contig}}\, \text{contig}_{ij} + \beta_{\text{lang}}\, \text{comlang}_{ij} + \beta_{\text{rta}}\, \text{rta}_{ij} + \varepsilon_{ij}, \qquad X_{ij} > 0. \tag{3}$$

Because $\ln X_{ij}$ is undefined when $X_{ij} = 0$, all zero-trade pairs are dropped. The coefficients are estimated by OLS with the exporter and importer fixed effects entered as dummy variables, and inference uses HC1 heteroskedasticity-robust standard errors.

### 4.1.2 PPML on the level of trade

PPML estimates equation (2) directly in levels. It is the pseudo-maximum-likelihood estimator associated with the Poisson family: the estimator solves the first-order (moment) conditions

$$\sum_{ij} \Big(X_{ij} - \exp(\mathbf{z}_{ij}'\boldsymbol{\beta})\Big)\mathbf{z}_{ij} = \mathbf{0}, \tag{4}$$

where $\mathbf{z}_{ij}$ collects the fixed-effect dummies and bilateral covariates. Crucially, equation (4) is well defined when $X_{ij} = 0$, so *zero trade flows are retained.* PPML does not require the data to be counts; it only requires the conditional mean to be correctly specified as in equation (2), and it is consistent under heteroskedasticity of essentially arbitrary form (Santos Silva and Tenreyro, 2006). We use a generalized-linear-model Poisson fit with HC1 robust (sandwich) standard errors, which are essential for valid PPML inference.

## 4.2 Exporter and importer fixed effects and the absorption of monadic GDPs

The user's specification calls for log exporter GDP and log importer GDP as covariates *and* for exporter and importer fixed effects. In a single cross-section these two requirements collide: the monadic GDP terms are perfectly collinear with the fixed effects and cannot be separately identified. The reason is mechanical. Exporter GDP, $\text{gdp}_i$, takes exactly one value for each exporter $i$; it does not vary across the importers $j$ that $i$ trades with. It is therefore constant *within* each exporter group, which makes $\ln \text{gdp}_i$ an exact linear combination of the exporter dummy variables $\{\mu_i\}$. Symmetrically, $\ln \text{gdp}_j$ is an exact linear combination of the importer dummies $\{\nu_j\}$. Any attempt to include both the monadic GDP and the full set of country dummies produces a rank-deficient design matrix; the GDP coefficients are not estimable and are absorbed by the fixed effects.

We verified this absorption directly and non-trivially. For every exporter we computed the within-group variance of $\ln \text{gdp}_i$, and for every importer the within-group variance of $\ln \text{gdp}_j$. The maximum of these within-group variances is exactly 0 in both samples (to machine precision), confirming that the monadic GDPs carry no within-country variation and are fully spanned by the fixed effects. This is not a limitation of the estimator but a feature of structural gravity: as Fally (2015) shows, the exporter and importer fixed effects are precisely the objects that capture each country's size and multilateral resistance, so the size variables are subsumed *by design*. The coefficients we report and interpret are therefore those on the genuinely bilateral (dyadic) covariates—distance, contiguity, common language, and RTA—which are identified from within-exporter and within-importer variation.

**Why monadic GDPs drop out**

With exporter and importer fixed effects in a single cross-section, $\ln \text{gdp}_i$ is constant within exporter $i$ and $\ln \text{gdp}_j$ is constant within importer $j$. Each is thus a perfect linear combination of the fixed-effect dummies (measured within-group variance $= 0$), so the GDP coefficients are not separately identified. This is the standard structural-gravity result of Fally (2015), not a data problem.

## 4.3 Excluding self-trade: a critical methodological decision

The single most consequential sample-construction decision in this analysis is the exclusion of *self-trade*—the diagonal observations for which the exporter equals the importer (`iso3_o` = `iso3_d`). The CEPII BACI trade flows measure *international* trade only; a country's trade with itself (internal or intra-national trade) is not observed and appears as missing. If these diagonal rows are retained, the PPML step fills the missing internal flow with a structural zero, which asserts—counterfactually—that each country ships nothing to itself. Worse, the recorded "distance" on the diagonal is a small internal distance, so these artificial zero-at-short-distance observations pull the fitted distance-trade relationship toward zero and severely bias the distance elasticity.

The magnitude of this bias is not academic. In an initial specification that retained the 267 self-trade rows, the PPML distance elasticity was an implausible −0.316; excluding self-trade restored it to −0.648, squarely within the range reported in the meta-analysis of Disdier and Head (2008) and in Head and Mayer (2014). Dropping internal trade when it is unobserved is standard practice in the gravity literature (Head and Mayer, 2014; Santos Silva and Tenreyro, 2006). Accordingly, both the OLS and PPML samples exclude the diagonal. Concretely, from the 63,504 rows we drop 503 fully-duplicated dyad rows to reach 63,001, then drop 267 self-trade rows to reach 62,734 international-pair rows, before applying covariate completeness.

## 4.4 Sample construction

The task's covariate requirement—"non-missing and positive"—is applied in the econometrically correct way. The continuous covariates that enter in logs (both GDPs and distance) are required to be non-missing and strictly positive. The binary dummies (contiguity, common language, RTA) are required only to be non-missing; requiring them to be positive would nonsensically discard every non-contiguous pair, every pair without a shared language, and every pair without an RTA. Applying these rules to the 62,734 international-pair rows yields 39,800 covariate-complete observations.

- **OLS sample.** The subset with `tradeflow_baci` $> 0$: 28,113 observations. The dependent variable is ln(trade).
- **PPML sample.** All 39,800 covariate-complete rows, with missing trade set to 0: 28,113 positive flows plus 11,687 structural zeros. The dependent variable is the trade level.

The exporter and importer fixed effects span 198 exporters and 198 importers in the OLS sample and 200 of each in the PPML sample (the larger set reflects countries that appear only in zero-flow pairs). Figure 2 visualizes the difference: PPML recovers the 11,687 zero-flow pairs that OLS is forced to discard.

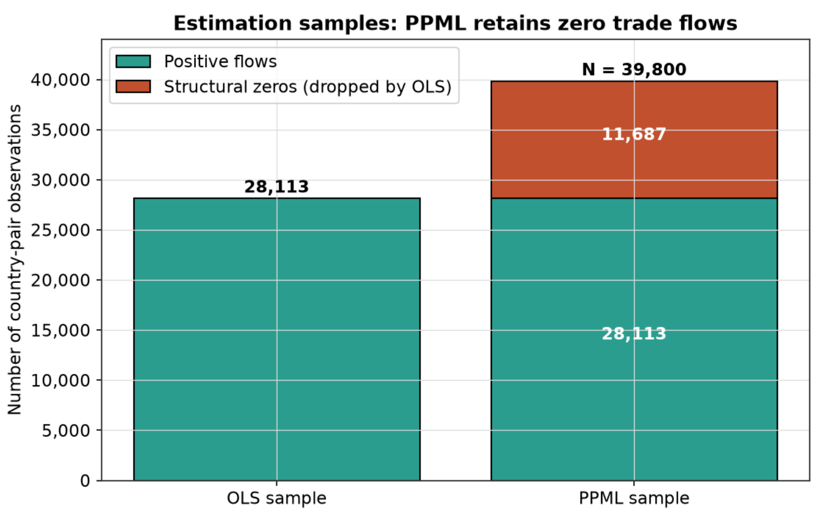



**Figure 2:** Estimation samples. OLS uses only the 28,113 positive flows; PPML additionally retains the 11,687 structural zeros, for $N = 39{,}800$. The zero flows constitute the extensive margin of trade that log-OLS cannot use.

## 4.5 Reproducibility

The analysis is fully reproducible. Data acquisition (`download_data.py`) streams the V202211 archive, filters to 2019, and writes `gravity_2019.csv`. Estimation (`estimate_gravity.py`) constructs the samples, verifies GDP absorption, fits both models, and writes the two regression tables and a JSON summary. The random seed is fixed at 42; estimation uses `statsmodels` 0.14.6 (Seabold and Perktold, 2010) with `pandas` 3.0.3 and `numpy` 2.5.1. The PPML GLM converged in fewer than 200 iterations. Section 8 reproduces the core estimation code.

# 5 Results

## 5.1 The raw gravity relationship

Before turning to the fitted models, Figure 3 shows the raw 2019 relationship between log trade and log distance across all positive-flow international pairs. The strong negative gradient—trade falls sharply as partners grow farther apart—is the empirical regularity the gravity model is built to explain, and it confirms that the estimated distance elasticities below are recovering a genuine feature of the data.

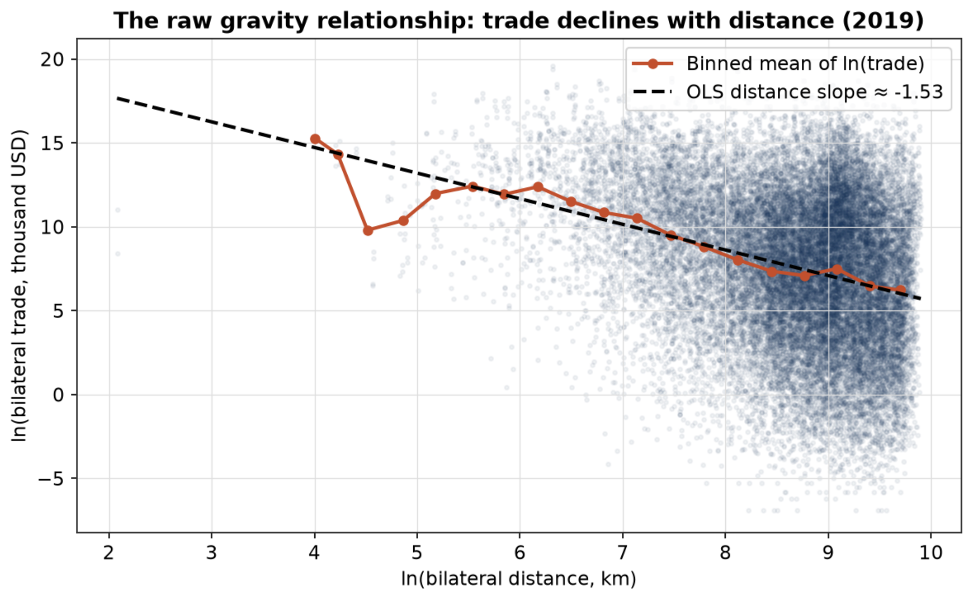



**Figure 3:** The raw gravity relationship in 2019. Each point is a positive-flow country pair; the red line traces the binned mean of log trade, and the dashed line shows the OLS distance slope. Trade declines monotonically with distance.

## 5.2 Regression tables

Tables 2 and 3 report the two sets of estimates. Both tables reproduce exactly the machine-generated CSV files `ols_regression_table.csv` and `ppml_regression_table.csv`. Coefficients on the bilateral covariates are shown with HC1 robust standard errors, test statistics, $p$-values, and 95% confidence intervals; the diagnostics block reports the sample size, the (pseudo-)$R^2$, the distance elasticity, and the RTA effect.

**Table 2:** OLS on log trade (positive flows only). Dependent variable: ln(tradeflow_baci). Exporter and importer fixed effects included (198 levels each). HC1 robust standard errors.

| Variable | Coef. | Robust SE | $t$ | $p$ | 95% CI |
|---|---|---|---|---|---|
| Log distance | −1.527,9 | 0.022,8 | −66.95 | 0.000 | [−1.573, −1.483] |
| Contiguity | 0.994,7 | 0.117,4 | 8.47 | 0.000 | [ 0.765, 1.225] |
| Common official language | 0.886,5 | 0.046,0 | 19.28 | 0.000 | [ 0.796, 0.977] |
| Regional trade agreement | 0.452,6 | 0.039,6 | 11.43 | 0.000 | [ 0.375, 0.530] |
| *Diagnostics* | | | | | |
| Observations ($N$) | 28,113 | | | | |
| $R^2$ | 0.7638 | | | | |
| Distance elasticity ($\beta_{\text{dist}}$) | −1.5279 | | | | |
| RTA trade effect $100(e^{\beta_{\text{rta}}} - 1)$ | +57.24% | | | | |
| Monadic GDPs (`ln_gdp_o`, `ln_gdp_d`) | Absorbed by fixed effects (perfect collinearity) | | | | |
| Fixed effects | Exporter (198) + Importer (198) | | | | |

**Table 3:** PPML on the level of trade (structural zeros retained). Dependent variable: `tradeflow_baci` in levels. Exporter and importer fixed effects included (200 levels each). HC1 robust standard errors.

| Variable | Coef. | Robust SE | $z$ | $p$ | 95% CI |
|---|---|---|---|---|---|
| Log distance | −0.648,2 | 0.026,3 | −24.65 | 0.000 | [−0.700, −0.597] |
| Contiguity | 0.554,3 | 0.072,3 | 7.67 | 0.000 | [ 0.413, 0.696] |
| Common official language | −0.001,6 | 0.069,1 | −0.02 | 0.981 | [−0.137, 0.134] |
| Regional trade agreement | 0.351,0 | 0.052,3 | 6.71 | 0.000 | [ 0.249, 0.453] |
| *Diagnostics* | | | | | |
| Observations ($N$) | 39,800 (28,113 positive + 11,687 zeros) | | | | |
| Pseudo-$R^2$ [corr$(Y, \hat{Y})^2$] | 0.8782 | | | | |
| Distance elasticity ($\beta_{\text{dist}}$) | −0.6482 | | | | |
| RTA trade effect $100(e^{\beta_{\text{rta}}} - 1)$ | +42.05% | | | | |
| Monadic GDPs (`ln_gdp_o`, `ln_gdp_d`) | Absorbed by fixed effects (perfect collinearity) | | | | |
| Fixed effects | Exporter (200) + Importer (200) | | | | |

## 5.3 The four requested quantities, by estimator

Table 4 places the four requested quantities side by side.

**Table 4:** Requested quantities by estimator (CEPII V202211, 2019).

| Quantity | OLS (log trade) | PPML (levels) |
|---|---|---|
| Distance elasticity ($\beta_{\text{dist}}$) | −1.528 | −0.648 |
| RTA coefficient ($\beta_{\text{rta}}$) | 0.453 | 0.351 |
| RTA % trade effect $100(e^{\beta} - 1)$ | +57.2% | +42.0% |
| Number of observations $N$ | 28,113 | 39,800 |
| (Pseudo-)$R^2$ | 0.764 ($R^2$) | 0.878 (pseudo-$R^2$) |

**Distance elasticity.** Both estimators recover the expected negative, highly significant distance elasticity, but their magnitudes diverge sharply. OLS returns −1.528: a 10% increase in distance is associated with a 15% fall in log trade. PPML returns −0.648: a 10% increase in distance lowers expected trade by about 6.5%. The PPML value sits comfortably within the empirical consensus range (roughly −0.6 to −1.1) synthesized by Disdier and Head (2008) and Head and Mayer (2014), whereas the OLS value is at the extreme high end, inflated by the log-linearization bias discussed in Section 6.

**RTA effect.** The RTA coefficient is positive and significant under both estimators. Converting to a percentage trade effect via $100(e^{\beta} - 1)$, OLS implies that RTA membership raises bilateral trade by 57.2%, whereas PPML implies a more modest 42.0%. The PPML figure is in line with the partial (direct) RTA effects reported in the trade-agreement literature (Baier and Bergstrand, 2007).

**Number of observations.** This is where the estimators differ most visibly. OLS is confined to the 28,113 positive flows. PPML retains the full covariate-complete sample of 39,800, adding the 11,687 zero-trade pairs that OLS must discard—an increase of about 42% in the estimation sample. These zeros are not missing at random; they are disproportionately small, distant, or economically dissimilar pairs, so discarding them (as OLS does) is a non-random truncation of the sample.

**Goodness of fit.** The OLS $R^2$ is 0.764; the PPML pseudo-$R^2$, defined as the squared correlation between actual and fitted trade *levels*, is 0.878. The two numbers are not directly comparable—one is computed in logs on positive flows, the other in levels on the full sample—but both indicate that the fixed-effects gravity specification explains the large majority of the variation in trade.

## 5.4 Visual comparison

Figures 4 and 5 summarize the coefficient estimates and the two headline quantities. The contrast in the distance coefficient is the visual centerpiece: the OLS point estimate lies far to the left of the PPML estimate, with non-overlapping confidence intervals.

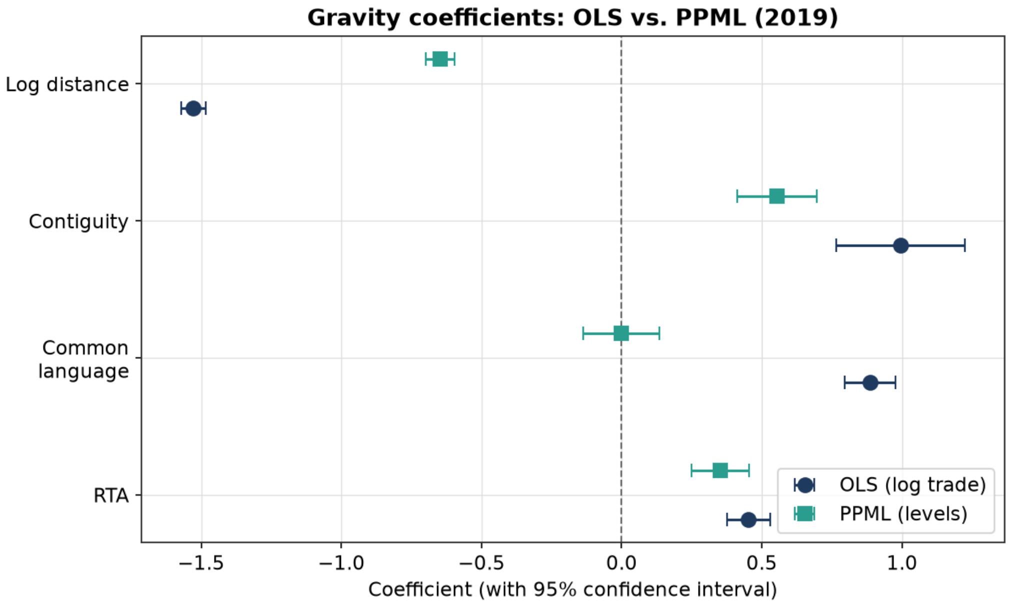



**Figure 4:** Coefficient estimates with 95% confidence intervals, OLS (circles) versus PPML (squares). The distance and contiguity coefficients are substantially larger in magnitude under OLS. Common language is significant under OLS but statistically indistinguishable from zero under PPML.

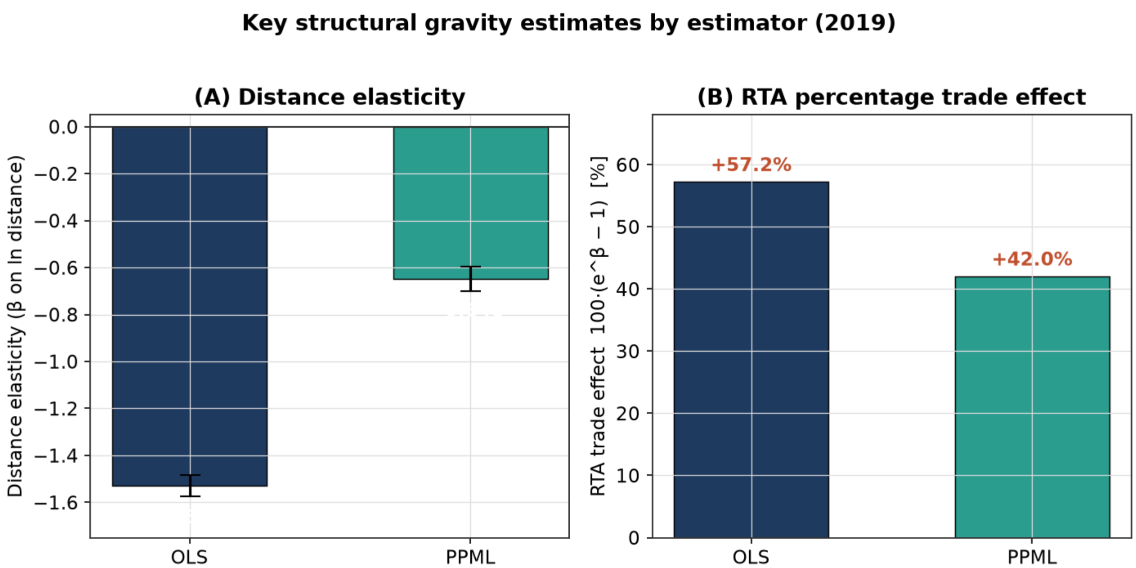



**Figure 5:** The two headline quantities. (A) Distance elasticity: OLS (−1.53) is roughly 2.4× larger in magnitude than PPML (−0.65). (B) RTA percentage trade effect: OLS (+57.2%) exceeds PPML (+42.0%) by about a third. Error bars in (A) are 95% confidence intervals.

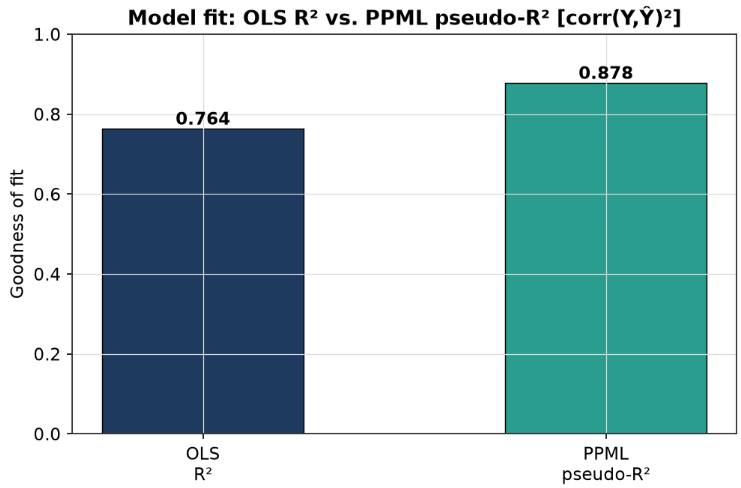



**Figure 6:** Goodness of fit. OLS $R^2$ (in logs, positive flows) versus PPML pseudo-$R^2$ defined as $\mathrm{corr}(Y, \hat{Y})^2$ in levels on the full sample.

# 6 Discussion

## 6.1 Why the estimator changes the answer

The two estimators fit the *same* structural specification to the *same* year of data, yet they deliver a distance elasticity that differs by more than a factor of two and an RTA effect that differs by about a third. The divergence is not sampling noise—the confidence intervals do not overlap—but the signature of two well-understood econometric problems that afflict log-OLS and that PPML is designed to avoid.

### 6.1.1 The consistent treatment of zero trade flows

The first problem is selection on the dependent variable. Roughly 29% of the covariate-complete international pairs in 2019 record zero trade (11,687 of 39,800). Because $\ln 0$ is undefined, log-OLS can only be estimated after these pairs are deleted, which restricts the sample to 28,113 observations. This deletion is not innocuous: zero-trade pairs are systematically different—they tend to be smaller, more distant, and less integrated—so dropping them truncates the sample precisely on the dimensions the gravity model is trying to measure. The result is a form of sample-selection bias in which the surviving positive flows over-represent well-connected pairs.

PPML, by contrast, is defined on the level of trade and its moment condition (4) is perfectly well behaved at $X_{ij} = 0$. It therefore uses the entire covariate-complete sample, including every zero. The zeros are informative: they identify the extensive margin (whether a pair trades at all), and retaining them is what allows PPML to recover a distance elasticity that reflects the full population of country pairs rather than the selected subset that happens to trade. This is the first-order reason PPML is preferred in modern structural gravity work (Santos Silva and Tenreyro, 2006, 2022; Yotov et al., 2016).

### 6.1.2 Heteroskedasticity and Jensen's inequality

The second problem is more subtle and would bias log-OLS *even if there were no zeros at all.* Trade data are severely heteroskedastic: the variance of trade flows grows with their level. The multiplicative gravity model is naturally written as $X_{ij} = \exp(\mathbf{z}'_{ij}\boldsymbol{\beta})\,\eta_{ij}$ with $\mathbb{E}[\eta_{ij} \mid \mathbf{z}_{ij}] = 1$. Taking logs gives $\ln X_{ij} = \mathbf{z}'_{ij}\boldsymbol{\beta} + \ln \eta_{ij}$, and OLS on this equation requires $\mathbb{E}[\ln \eta_{ij} \mid \mathbf{z}_{ij}]$ to be constant. But by Jensen's inequality, for a non-degenerate positive random variable,

$$\mathbb{E}[\ln \eta_{ij} \mid \mathbf{z}_{ij}] \;\neq\; \ln \mathbb{E}[\eta_{ij} \mid \mathbf{z}_{ij}] \;=\; \ln 1 = 0, \tag{5}$$

and when the error is heteroskedastic—so that the spread of $\eta_{ij}$ depends on the covariates $\mathbf{z}_{ij}$—the conditional mean $\mathbb{E}[\ln \eta_{ij} \mid \mathbf{z}_{ij}]$ is itself a function of the covariates. That function is then absorbed into the estimated slopes, biasing them. Santos Silva and Tenreyro (2006) showed formally that this makes log-OLS inconsistent for the parameters of the conditional mean, and that the sign and size of the bias depend on the (unknown) pattern of heteroskedasticity.

PPML sidesteps the problem entirely because it never takes logs of the dependent variable. It estimates the conditional mean $\mathbb{E}[X_{ij} \mid \mathbf{z}_{ij}] = \exp(\mathbf{z}'_{ij}\boldsymbol{\beta})$ directly, and its moment condition (4) is valid under arbitrary heteroskedasticity, delivering consistent estimates provided only that the conditional mean is correctly specified. This is why the PPML distance elasticity ($-0.648$) is the credible one and the OLS elasticity ($-1.528$) is inflated: the log transformation has folded the distance-dependent error variance into the distance slope.

**Two reasons PPML wins**

**(1) Zeros.** PPML is defined in levels, so it keeps the 11,687 zero-trade pairs that OLS must delete, avoiding selection on the dependent variable. **(2) Heteroskedasticity / Jensen.** PPML models $\mathbb{E}[X] = \exp(\cdot)$ directly, so it is consistent under arbitrary heteroskedasticity, whereas log-OLS confounds the distance-dependent error variance with the distance slope because $\mathbb{E}[\ln X] \neq \ln \mathbb{E}[X]$.

### 6.2 The self-trade correction

The distance elasticity was also the diagnostic that exposed a data problem. An early PPML run that inadvertently retained the 267 self-trade (diagonal) rows produced a distance elasticity of only $-0.316$—far too small relative to the literature. The cause was mechanical: BACI does not observe internal trade, so the diagonal flows were filled with zeros at very short internal distances, injecting a cluster of "zero trade at near-zero distance" points that flattened the estimated distance gradient. Excluding self-trade—standard practice when internal trade is unavailable (Head and Mayer, 2014; Santos Silva and Tenreyro, 2006)—moved the elasticity to $-0.648$, a value fully consistent with Disdier and Head (2008). The episode is a reminder that in gravity estimation the treatment of the sample (which observations are trade, which are zeros, which are simply unobserved) can matter as much as the choice of estimator.

### 6.3 Interpreting the individual covariates

Beyond distance, the two estimators tell a broadly consistent but quantitatively different story. Contiguity and RTA membership are positive and significant under both, though smaller under PPML—again reflecting the removal of the log-linearization bias. The most striking difference is common official language, which is large and highly significant under OLS ($\beta = 0.887$) but statistically indistinguishable from zero under PPML ($\beta = -0.002$, $p = 0.98$). This pattern is itself informative: much of the apparent "language effect" in log-OLS operates through which pairs trade at all (the extensive margin) and through heteroskedasticity, both of which PPML handles differently. Once zeros are retained and the conditional mean is modeled directly, the marginal contribution of a shared official language—conditional on exporter and importer fixed effects, distance, contiguity, and RTA membership—is negligible in this single cross-section. This does not overturn the broader evidence that language lowers trade costs; it shows that, in this specification and year, the effect is not separately identified once the other frictions and the extensive margin are properly accounted for.

### 6.4 Limitations

Three caveats bound the interpretation of these estimates. First, this is a single 2019 cross-section, so the fixed effects absorb multilateral resistance but the analysis cannot exploit the within-pair, over-time variation that panel gravity uses to identify RTA effects free of endogeneity (Baier and Bergstrand, 2007); the RTA coefficients here are best read as associations, not causal treatment effects. Second, standard errors are HC1 heteroskedasticity-robust but not clustered; multi-way clustering at the exporter and importer level would widen the intervals and is a natural robustness extension. Third, because internal trade is unobserved, the analysis is restricted to international flows; a specification that incorporated intra-national trade (as in the most recent structural-gravity practice) could sharpen the distance elasticity further. None of these caveats affects the central comparison, which is about the estimator, not the data year.

## 7 Conclusion

Fitting the same structural gravity model to the 2019 cross-section of the CEPII Gravity Database (V202211) two ways yields two clear lessons. *Empirically*, the gravity relationship is strong and in the expected direction under both estimators: trade falls with distance and rises with contiguity and RTA membership. But the magnitudes depend heavily on the estimator. OLS on log trade delivers a distance elasticity of $-1.528$ and an RTA effect of $+57.2\%$ on 28,113 positive flows ($R^2 = 0.764$); PPML on the level of trade delivers a distance elasticity of $-0.648$ and an RTA effect of $+42.0\%$ on 39,800 observations that retain 11,687 zeros (pseudo-$R^2 = 0.878$). The OLS elasticity is inflated by roughly a factor of 2.4.

*Methodologically*, PPML is the preferred estimator, for the two reasons that the modern trade literature has converged upon: it treats zero trade flows consistently rather than deleting them, and it is robust to the heteroskedasticity that makes log-OLS inconsistent through Jensen's inequality. The analysis also underscores that sound sample construction is a precondition for credible estimates: excluding unobserved self-trade moved the PPML distance elasticity from an implausible $-0.316$ to a literature-consistent $-0.648$. Finally, the exporter and importer fixed effects that make the specification structural also absorb the monadic GDPs by construction—a feature of the model, not a defect of the data. Taken together, the results are a compact demonstration of why PPML with exporter and importer fixed effects has become the workhorse of empirical trade analysis.

## 8 Reproducible Code

The complete estimation routine is reproduced below (the full acquisition and estimation scripts, `download_data.py` and `estimate_gravity.py`, accompany this report). It builds both samples, verifies the absorption of the monadic GDPs, fits OLS and PPML, and writes the two regression tables.

```python
import numpy as np, pandas as pd
import statsmodels.api as sm
import statsmodels.formula.api as smf

df = pd.read_csv("data/gravity_2019.csv").drop_duplicates()
```

```python

# Exclude self-trade (diagonal): BACI observes international trade only.
df = df[df["iso3_o"] != df["iso3_d"]].copy()

# Covariate completeness: continuous vars positive; dummies non-missing.
m = (df[["gdp_o", "gdp_d", "dist"]].gt(0).all(axis=1)
     & df[["gdp_o", "gdp_d", "dist", "contig", "comlang_off", "rta"]].notna().all(axis=1))
base = df[m].copy()
base["ln_gdp_o"] = np.log(base["gdp_o"])
base["ln_gdp_d"] = np.log(base["gdp_d"])
base["ln_dist"]  = np.log(base["dist"])

# OLS sample: positive flows only (log excludes zeros).
ols_df = base[base["tradeflow_baci"] > 0].copy()
ols_df["ln_trade"] = np.log(ols_df["tradeflow_baci"])

# PPML sample: retain structural zeros (missing trade -> 0).
ppml_df = base.copy()
ppml_df["trade_level"] = ppml_df["tradeflow_baci"].fillna(0.0).clip(lower=0.0)

# Monadic GDP absorption check: within-group variance == 0.
assert base.groupby("iso3_o")["ln_gdp_o"].var(ddof=0).max() < 1e-12
assert base.groupby("iso3_d")["ln_gdp_d"].var(ddof=0).max() < 1e-12

rhs = "ln_dist + contig + comlang_off + rta + C(iso3_o) + C(iso3_d)"

# OLS on log trade, HC1 robust SE.
ols = smf.ols(f"ln_trade ~ {rhs}", data=ols_df).fit(cov_type="HC1")

# PPML on trade level, HC1 robust SE.
ppml = smf.glm(f"trade_level ~ {rhs}", data=ppml_df,
               family=sm.families.Poisson()).fit(cov_type="HC1", maxiter=200)

# RTA percentage trade effect.
for r in (ols, ppml):
    print("dist elasticity:", r.params["ln_dist"],
          "| RTA %:", 100 * (np.exp(r.params["rta"]) - 1))
```

**Listing 1:** Core of `estimate_gravity.py`: sample construction, GDP-absorption check, and the two estimators.

# References


Anderson, J. E. and van Wincoop, E. (2003). Gravity with gravitas: A solution to the border puzzle. *American Economic Review*, 93(1):170–192.

Baier, S. L. and Bergstrand, J. H. (2007). Do free trade agreements actually increase members' international trade? *Journal of International Economics*, 71(1):72–95.

Conte, M., Cotterlaz, P., and Mayer, T. (2022). The cepii gravity database. CEPII Working Paper 2022-05, CEPII.

Correia, S., Guimarães, P., and Zylkin, T. (2020). Fast poisson estimation with high-dimensional fixed effects. *The Stata Journal*, 20(1):95–115.

Disdier, A.-C. and Head, K. (2008). The puzzling persistence of the distance effect on bilateral trade. *The Review of Economics and Statistics*, 90(1):37–48.

Fally, T. (2015). Structural gravity and fixed effects. *Journal of International Economics*, 97(1):76–85.

Gaulier, G. and Zignago, S. (2010). Baci: International trade database at the product-level. the 1994–2007 version. CEPII Working Paper 2010-23, CEPII.

Head, K. and Mayer, T. (2014). Gravity equations: Workhorse, toolkit, and cookbook. In Gopinath, G., Helpman, E., and Rogoff, K., editors, *Handbook of International Economics*, volume 4, pages 131–195. Elsevier.

Santos Silva, J. M. C. and Tenreyro, S. (2006). The log of gravity. *The Review of Economics and Statistics*, 88(4):641–658.

Santos Silva, J. M. C. and Tenreyro, S. (2022). The log of gravity at 15. *Portuguese Economic Journal*, 21(1):1–24.

Seabold, S. and Perktold, J. (2010). statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, pages 92–96.

Tinbergen, J. (1962). *Shaping the World Economy: Suggestions for an International Economic Policy*. Twentieth Century Fund, New York.

Yotov, Y. V., Piermartini, R., Monteiro, J.-A., and Larch, M. (2016). *An Advanced Guide to Trade Policy Analysis: The Structural Gravity Model*. World Trade Organization and United Nations, Geneva.