Adam, Hanna L.; Larch, Mario; Nower, Michael

Working Paper

ANOVA-HDFE: Fast Variance Decomposition with High-Dimensional Fixed Effects and an Application to Trade Flows

CESifo Working Paper, No. 12055

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Adam, Hanna L.; Larch, Mario; Nower, Michael (2025) : ANOVA-HDFE: Fast Variance Decomposition with High-Dimensional Fixed Effects and an Application to Trade Flows, CESifo Working Paper, No. 12055, Munich Society for the Promotion of Economic Research - CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/327665

CES ifo Working Papers

12055
2025

August 2025

ANOVA-HDFE: Fast Variance Decomposition with High-Dimensional Fixed Effects and an Application to Trade Flows

Hanna L. Adam, Mario Larch, Michael Nower

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

ANOVA-HDFE: Fast Variance Decomposition with High-Dimensional Fixed Effects and an Application to Trade Flows

Hanna L. Adam
University of Bayreuth

Mario Larch
University of Bayreuth
CEPII, ifo, CESifo, WIFO, GEP

Michael Nower[‡]
Durham University

August 5, 2025

Abstract

Performing an analysis of variance (ANOVA) on a large dataset spanning many dimensions becomes computationally challenging or even infeasible. We develop a new, fast procedure, ANOVA-HDFE, which uses sequential linear regressions and builds on recent advances in regression analysis with high-dimensional fixed effects (HDFE). It accommodates both balanced and unbalanced settings with many categorical and continuous covariates, while also allowing for high-dimensional fixed effects. Applying ANOVA-HDFE to bilateral trade flows, we find that 60% of the variation is at the country or country-time level. Moreover, a substantial proportion of the pair-specific variation remains unexplained by standard trade cost proxy variables.

JEL Classification Codes: F14, C23, C55, F16.
Keywords: Analysis of Variance, High-Dimensional Fixed Effects, Large Data, Variation in High Dimensions, Variation of Bilateral Trade Flows, Asymmetric Trade Costs, ANOVA-HDFE.

[*]Contact information: Adam—Department of Law, Business Administration & Economics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany; E-mail: hanna.adam@uni-bayreuth.de; Larch—Department of Law, Business Administration & Economics, University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany; Centre d'Etudes Prospectives et d'Informations Internationales (CEPII), Paris, France; ifo Institute and CESifo Research Network, Munich, Germany; Austrian Institute of Economic Research (WIFO), Vienna, Austria; Nottingham Centre for Research on Globalisation and Economic Policy (GEP), Nottingham, United Kingdom. E-mail: mario.larch@uni-bayreuth.de; Nower—Department of Economics, Durham University; E-mail: michael.nower@durham.ac.uk.

[†]Acknowledgements to be added later.

1 Introduction

In recent years, research data in economics and many other fields have become larger and cover more dimensions, such as data on individuals, firms, workers, or countries, that are observed over years and occupations or industries. A key concern for researchers using datasets with many dimensions is understanding the strength of the sources of variation in an outcome variable of interest across the different dimensions and the extent to which other available covariates contribute to explaining this variation. The most commonly used method to disentangle the contribution of different factors to an overall outcome is an analysis of variance (ANOVA), which decomposes the variation of the outcome variable of interest into components accounted for by different factors. Standard ANOVA implementation procedures struggle to provide (fast) results for large datasets spanning many dimensions.

In advance of ANOVA, regression analysis of large datasets has increasingly used high-dimensional fixed effects to control for a wide array of unobservable factors across many dimensions of variation in a variety of fields, including labor markets (Torres et al., 2018; Mittag, 2019), climate and ecological economics (Zhu et al., 2021; Zhang et al., 2024), aid (Rommel and Schaudt, 2020), and trade (Larch et al., 2019; Borchert et al., 2022; Larch and Yotov, 2024), to name just a few. This increased usage of high-dimensional fixed effects has been facilitated by the development of statistical packages for fast regression analysis with high-dimensional fixed effects (see Guimarães and Portugal, 2010; Gaure, 2013b,a; Correia, 2016; Cornelissen, 2018).

Building on these developments in regression analysis, we develop a novel procedure, ANOVA-HDFE, for feasible and fast ANOVA with high-dimensional fixed effects using sequential linear regressions. ANOVA-HDFE allows one to perform an ANOVA with a combination of continuous and categorical covariates, alongside rich sets of fixed effects, in both balanced and unbalanced settings. A key benefit, above performing ANOVA for data with any (high) dimension of fixed effects, is that ANOVA-HDFE offers substantial speed improvements in the presence of many explanatory variables, regardless of the number of dimensions. Hence, practitioners can obtain a first quick impression of the variation

in their outcome variable of interest and the potential influence of different explanatory variables by applying ANOVA-HDFE before proceeding with a more in-depth analysis.

To illustrate our new procedure and its benefits, we apply it to decompose the variance of bilateral international trade flows, similar to Egger and Nigai (2015), Cuñat and Zymek (2024), Redding and Weinstein (2024), and Gervais (2025), but with a broader scope, including more countries and years. Alongside demonstrating the computational speed benefits of ANOVA-HDFE, our application to an extensive trade dataset with many countries and years highlights two key insights. First, about 60% of the variation in bilateral trade flows is at the country or country-time level, i.e. is accounted for by the exporter- and importer-(time) dimensions, while about 20% of the remaining bilateral variation is asymmetric. Second, we reveal that a substantial proportion of the pair-specific variation in trade flows cannot be explained by standard trade cost proxy variables (such as distance or Regional Trade Agreements—RTAs), which are typically used in empirical gravity analyses of trade.

The remainder of the paper is organized as follows. In the following section, we introduce ANOVA with high-dimensional fixed effects, provide formulas for ANOVA of a strongly balanced directional three-way panel, and discuss the standard ANOVA and our fast ANOVA-HDFE implementations in statistical software. In Section 3, we present the data used to demonstrate the ANOVA implementations. Section 4 discusses the results for a variance decomposition of bilateral trade flows along various dimensions, including speed comparisons between the standard and our new implementations. Section 5 demonstrates the use of ANOVA-HDFE with covariates (i.e. analysis of covariance) by providing a variance analysis of the bilateral variation in trade flows, including various trade cost proxy variables. The last section concludes.

2 Theory and Method: Analysis of Variance

An ANOVA decomposes the variation of a variable of interest, measured in sums of squares, into components accounted for by different explanatory variables (see Fisher, 1925; Scheffé, 1999; Sahai and Ageel, 2000; Searle et al., 2006). The sequential sums of

squares method splits up the entire variation of the variable of interest into the different components, such that the sums of squares of the components add up to the total sum of squares. Hence, the sequential sums of squares assess how much of the variable's variation is explained by sequentially adding variables to a linear regression model. In the case where the explanatory variables share some overlapping variation, the order in which the variables are added affects their respective explanatory power. The researcher has to choose the order of the variables. Theory on the overlapping nature of some dimensions may provide a guide to the most suitable order. For example, international trade flows vary by exporter, by importer, and by time. Although there is no guidance on whether to add the time or country dimension first, for assessing the variation explained by the exporter dimension and the variation explained by the exporter-time dimension, one needs to add the exporter variables first and the exporter-time variables later to see what variation beyond the exporter variation is explained by the exporter-time dimension. If one first added the exporter-time variables, there would be no variation remaining to be explained by the simple exporter dimension. In plant-level labor market analyses, where outcomes vary by firm, plant, worker, and time, one would add firm variables before the plant and worker variables, as each plant belongs to one firm, and each worker belongs to one plant (assuming that there are no switchers). For similar arguments as for trade, the interactions of the variables have to be added afterwards.

Although ANOVA is based on a linear model, it differs from ordinary least squares (OLS), which is typically used to fit linear models. While OLS fits linear models where perfectly overlapping dimensions cannot be disentangled, ANOVA can disentangle perfectly overlapping dimensions, such as in the above-mentioned example of the exporter and exporter-time dimensions in trade analyses, making ANOVA a valuable method.

2.1 ANOVA Theory in Balanced Three-Way Panels

We theoretically outline the decomposition of a variable whose variation comes from three dimensions, as commonly encountered, for example, in panel data of bilateral international

trade or migration flows, where the dimensions are origin, destination, and time.[1] In addition to the time dimension, observations are directional. Although it is possible to extend the theoretical definitions in further dimensions, we focus on directional three-dimensional variation in this section, matching our later application. Our new ANOVA-HDFE implementation procedure presented in Section 2.3 can be equally applied to any dimensions of variation.

Consider a strongly balanced panel of a variable of interest, denoted X_{ijt}, with T time periods ($t = 1, ...T$), I origins ($i = 1, ...I$), and J destinations ($j = 1, ...J$). The set of all origins is the same as the set of all destinations, however, they are distinguished, as observations are directional, that is, an observation from i to j in period t is typically not the same as an observation from j to i in period t.[2] We use the term "strongly balanced" panel to refer to the fact that observations are complete in all three dimensions. That is, for every time period, there is a directional observation from each origin to each destination. The panel also contains within-origin-destination observations, which are observations from a unit of origin to the same unit as destination in every time period. In total, there are $2T$ directional observations per pair ij and T within-origin-destination observations for each origin (or, equivalently, destination).

The total sum of squares of the three-way variable of interest, SQ_{total}, is the sum of squared deviations of the variable's observations from the grand mean $\overline{X}_{...}$, that is, from the mean over all observations of X_{ijt}:

$$SQ_{total} = \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\left(X_{ijt} - \overline{X}_{...}\right)^2. \quad (1)$$

SQ_{total} can be partitioned into the following components:

$$SQ_{total} = SQ_i + SQ_j + SQ_t + SQ_{it} + SQ_{jt} + SQ_{\overleftrightarrow{ij}} + SQ_{ij} + SQ_{\overleftrightarrow{ij}t} + SQ_{ijt}, \quad (2)$$

[1]For this exposition of partitioning trade flows into the sources of variation, we follow existing textbooks, such as Sahai and Ageel (2000); Searle et al. (2006). Our extension is the application to a directional three-way data structure.

[2]For our exposition, we refer to the three sources of variation, i, j, and t, as origin, destination, and time. However, other sources of variation may be applicable.

with

$$
\begin{aligned}
SQ_i &= JT \sum_{i=1}^{I} \left(\overline{X}_{i\cdot\cdot} - \overline{X}_{\cdots}\right)^2, \\
SQ_j &= IT \sum_{j=1}^{J} \left(\overline{X}_{\cdot j\cdot} - \overline{X}_{\cdots}\right)^2, \\
SQ_t &= IJ \sum_{t=1}^{T} \left(\overline{X}_{\cdot\cdot t} - \overline{X}_{\cdots}\right)^2, \\
SQ_{it} &= J \sum_{i=1}^{I} \sum_{t=1}^{T} \left(\overline{X}_{i\cdot t} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot\cdot t} + \overline{X}_{\cdots}\right)^2, \\
SQ_{jt} &= I \sum_{j=1}^{J} \sum_{t=1}^{T} \left(\overline{X}_{\cdot jt} - \overline{X}_{\cdot j\cdot} - \overline{X}_{\cdot\cdot t} + \overline{X}_{\cdots}\right)^2, \\
SQ_{\overleftrightarrow{ij}} &= T \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{X}_{\overleftrightarrow{ij}\cdot} - \overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot} - \overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot} + \overline{X}_{\cdots}\right)^2, \\
SQ_{ij} &= T \sum_{i=1}^{I} \sum_{j=1}^{J} \left(\overline{X}_{ij\cdot} - \overline{X}_{i\cdot\cdot} - \overline{X}_{\cdot j\cdot} - \overline{X}_{\overleftrightarrow{ij}\cdot} + \overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot} + \overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\right)^2, \\
SQ_{\overleftrightarrow{ij}t} &= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \Bigg(\overline{X}_{\overleftrightarrow{ij}t} - \overline{X}_{\cdots} - \overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t} - \overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t} + \overline{X}_{\cdot\cdot t} - \overline{X}_{\overleftrightarrow{ij}\cdot} \\
&\quad + \overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot} + \overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\Bigg)^2, \\
SQ_{ijt} &= \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{t=1}^{T} \Bigg(X_{ijt} + \overline{X}_{i\cdot\cdot} + \overline{X}_{\cdot j\cdot} - \overline{X}_{i\cdot t} - \overline{X}_{\cdot jt} + \overline{X}_{\overleftrightarrow{ij}\cdot} \\
&\quad - \overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot} - \overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot} - \overline{X}_{ij\cdot} - \overline{X}_{\overleftrightarrow{ij}t} + \overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t} + \overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t}\Bigg)^2,
\end{aligned}
$$

where a bar above a variable denotes the mean with respect to the given subindex. The subindex $\overleftrightarrow{ij}$ denotes the "symmetric" pair and includes observations from i to j as well as from j to i. More details on the derivation and definitions of the means can be found in Appendix A. It is possible to derive similar definitions for non-directional three-way settings, for example, strongly balanced firm, worker, and time data, or for settings with further dimensions.

In most data settings with several dimensions, there likely is variation at all levels, including their interactions, making the partitioning of variance across all dimensions relevant. However, under certain conditions, the above definitions no longer apply, making

them impractical for many applications. This is the case for unbalanced panels, when the number of observations is not complete for each dimension of variation. Unbalancedness can occur with data quality issues, where data is missing for some years, or simply due to an unbalanced existence. There are possible adjustments to make in unbalanced settings, depending on the reason why there are different numbers of observations per dimension (see, e.g., Sahai and Ageel, 2000, Chapters 4.10 and 5.13; Kendall et al., 1983, Chapter 35). However, these adjustments are not suitable for numerous applications. The above definitions also do not apply when assessing the explanatory power of further covariates taken from data, which do not simply span the dimensions of variation, such as indicator variables with different group sizes. Finally, if the covariates are continuous and are to be treated quantitatively, an analysis of covariance needs to be conducted. Theoretical treatment of analysis of covariance techniques is complicated (see Sahai and Ageel, 2000, Appendix N; mathematical details can be found in Scheffé, 1999, Chapter 6) and therefore impractical for application.

In applied settings, an often-used alternative to relying on analytical formulas is to compute ANOVAs numerically using statistical software. The computation of ANOVAs in statistical software is usually pre-packaged, allowing for a more flexible and user-friendly computation than manual mathematical partition, especially in unbalanced settings or when including further covariates (indicator or continuous), but also in strongly balanced cases where such partition is possible. Next, we illustrate the method implemented by one widely used software and outline some of its limitations before presenting our new implementation procedure.

2.2 ANOVA Method Implemented in Statistical Software

In this section, we illustrate how ANOVAs are implemented in statistical software, focusing on the software R. We demonstrate why computing ANOVAs for large datasets with many dimensions using pre-implemented commands can be computationally demanding. We briefly explain how the method is implemented in R's `aov` function used for fitting ANOVA models, closely following the exposition in Chambers and Hastie, 1992, Chapters

4 and 5; details can be found in Appendix B.

In R's `aov` function, sums of squares are computed as squared *effects*. The effects result from a QR decomposition that is routinely performed when estimating a linear model. All variables for which the amount of variation is to be computed are passed to the function, which can be indicator or continuous variables. The dimensions of variation are defined as fixed effects, that is, as an indicator variable for each unit within the respective dimension of variation. In the orthogonal transformation underlying the QR decomposition, the matrix of regressors is decomposed into a set of orthogonal vectors that span its columns. The effects are the product of the matrix of orthogonal vectors and the dependent variable vector.

The method using QR decomposition in R can be applied to settings with different dimensions of variation and also works in the unbalanced case, where application of the definitions outlined in Section 2.1 fails. It also allows including other covariates than those that define the dimensions of variation, which may be indicator variables (with potentially unequal group sizes) or continuous variables. Hence, without the need for any adjustment, it is possible to assess the explanatory power in a linear model for any variable.[3] However, the QR decomposition requires spanning the full set of variables in the matrix of regressors, which means that, when the dimensions of variation are to be captured by fixed effects, an indicator variable for each fixed effect is needed. Even if a variable is defined as a factor variable, containing all the fixed-effects levels, the full set of indicator variables is spanned in the computation. This makes computing ANOVAs for large datasets with variation in many dimensions computationally demanding and requires long running times.

[3]As explained above, in the sequential sums of squares, the variable ordering needs to be taken into account. That is, if the researcher is interested in assessing the explanatory power of a variable with variation in i, it needs to be included in the model before the fixed effects that capture the entire dimension of variation in i.

2.3 Implementation of ANOVA in Sequential Linear Regressions: ANOVA-HDFE

The above description shows that the computation of ANOVAs in R can be demanding in terms of computer capacity, due to the underlying QR decomposition. Considering the definition of sequential sums of squares, an ANOVA can also be performed by sequential linear regressions. In a first step, the variable of interest is regressed on the first explanatory variable, the R-squared is computed, and the residuals are extracted. The next step regresses these residuals on the first and the next variable, computes the R-squared, and, again, extracts the residuals. These residuals are the dependent variable of the next step, and so on. The values of the R-squared are then the fraction of the total sum of squares, or of the remainder of the total sum of squares, which is explained by the respective variable. Together with the total sum of squares calculated as in Equation (1), the sum of squares explained by each variable can be calculated. By sequentially adding the variables to the linear regression model, the residuals of one step are orthogonal to all the variables included previously and at the respective step. This orthogonalization between the included regressors and the residuals, undertaken by construction in OLS regression, draws a parallel to the QR decomposition used for common ANOVA commands, as the `aov` function in R described above, which involves computing a set of orthogonal vectors spanning the columns of the regressor matrix.

We develop our ANOVA-HDFE procedure for implementing ANOVAs using sequential linear regressions and relying on recent developments in estimating high-dimensional fixed effects models. These recent developments utilize the iterated weighted least squares approach (see, for example, Correia, 2016). Our implementation uses the `fixest` package from Bergé (2018), which offers fast estimation of econometric models with multiple fixed effects. Relying on fast high-dimensional fixed effects estimation methods and avoiding the spanning of the entire matrix of regressors in a QR decomposition saves time, particularly at later iterations, where many variables have already been added to the model.[4] Our

[4]Note that it is possible to include sets of variables at once, as long as the variables in the set are orthogonal to each other or interest lies in assessing the total explained variation of this set. For example, when considering trade data, different indicator variables, such as specific fixed effects or individual RTAs,

procedure considerably improves the computing time of ANOVAs in R, which we illustrate in Section 4, where we show results of applying our procedure to international trade data.[5] Although we illustrate our procedure in the setting of international trade flows, with directional three-way variation in exporter, importer, and time, it can equally be applied to settings with non-directional three-way variation, such as flows of international aid or firm-level labor market analyses. Its flexibility also allows application to settings with variation in less than or more than three dimensions—in either nested fashion, such as plant-level labor analyses within firms in panel settings, with variation in the firm, plant, worker, and time dimensions, where the plant variation (and, hence, plant fixed effects) are nested within the firm-level variation (and, hence, firm fixed effects)—or in non-nested fashion, such as product-level supply chain analyses with variation in the sending firm, receiving firm, product, and time dimensions.

3 Data

To apply our new procedure for computing ANOVAs with high-dimensional fixed effects to an international trade setting, we use trade flows in the period 1980–2016 from the WTO's Structural Gravity Manufacturing Database by Larch et al. (2025). The dataset contains bilateral international trade and domestic trade of manufactured goods for 223 countries, including all income levels.[6]

For our baseline ANOVA implementation, data on international and domestic trade are sufficient. Going beyond simply splitting the dimensions of variation, we utilize our new procedure to explore the extent to which the variables that are traditionally used in gravity analyses of trade can explain bilateral trade variation. This extension also demonstrates the flexibility of our procedure, illustrating its ability to combine high-dimensional fixed effects with indicator and continuous covariates in an ANOVA. For this application, we require data on standard trade cost variables, specifically bilateral distance, contiguity,

can be included at once. Doing so decreases the computation time substantially due to high-dimensional fixed effects procedures.

[5]The R-command `anovahdfe` and the replication package will be made available following publication.

[6]We drop Belgium-Luxembourg (BLX) due to missing control variables.

common language, colonial linkages, RTAs, WTO membership, and EU membership, all of which are obtained from the USITC's Dynamic Gravity Database. For RTAs, given our use of manufacturing trade data, we focus on bilateral preferential goods agreements notified to the WTO.

4 ANOVA-HDFE With Fixed Effects: Investigating the Dimensions of Variation of Bilateral Trade Flows

To illustrate our ANOVA-HDFE procedure, we analyze bilateral international trade flows, an example of a variable with directional three-way variation. Traded goods are shipped from an exporting country (origin dimension i) to an importing country (destination dimension j) in a given time period (time dimension t, in years). Our procedure allows us to assess the proportion of variation of trade flows (in logarithms) explained by different groups of fixed effects—which define the different dimensions of variation (i, j, t, and their combinations).[7] Specifically, we assess how much of the variation is country-specific (that is, explained by the exporter or importer fixed effects) or country-time-specific and how much of the variation is bilateral symmetric-(time-) or asymmetric-(time-)specific. Distinguishing the contribution of all these dimensions in explaining the variation of bilateral trade flows is not possible using regressions. For example, linear regression analysis would not allow including both exporter and exporter-time fixed effects due to collinearity. Including exporter fixed effects in our ANOVA before exporter-time fixed effects, however, allows us to quantify how much exporter-time variation there is beyond the exporter dimension.

With datasets of bilateral trade flows, it is only rarely possible to rely on the analytical equations of ANOVAs outlined in Section 2.1. Panel data on bilateral trade flows are often unbalanced, so that there are no longer exactly $2T$ observations of directional trade flows per country pair. For example, non-zero trade flows between two countries might only be observed in one direction in a given year. Additionally, the inclusion of trade cost

[7] We take the natural logarithm to log-linearize the gravity equation of international trade.

proxy variables (such as the distance between two countries or whether two countries have signed a trade agreement), which are continuous or lead to unbalanced groups, also precludes the simple application of the analytical equations.

We apply our ANOVA-HDFE procedure of sequential linear regressions described in Section 2.3 in R, relying on fast fixed effects estimation methods. To illustrate the improvement in the computing times of our implementation in R, we compare the computing times with those for ANOVA calculations using the standard `aov` function in R. Table 1 shows the computing times for both methods, with all the dimensions of variation outlined in Equation (2), for different samples.[8] The samples are different subsamples of our entire unbalanced dataset of international trade flows. The first column in Table 1 describes the sample and gives the maximum number of fixed effects (FE) that are generated for the respective sample and passed on to the functions.[9] Our ANOVA-HDFE implementation substantially decreases the computing time, being between 550 and approximately 2 million times faster than the `aov` function in delivering the same results. For the entire unbalanced panel dataset, the `aov` function fails to compute results due to memory constraints, while ANOVA-HDFE delivers results within 7 seconds.[10]

The different subsamples were selected to examine different features of the data and illustrate the ability of our ANOVA-HDFE procedure to perform ANOVA across different data characteristics. These subsamples include different levels of balancedness, as well as cross-sections and panels of different sizes. First, to rule out that the variation we are aiming to explain is influenced by whether or not the trade flows data is balanced, we run ANOVAs for different balanced subsamples.[11] Strongly balanced subsamples only include country pairs for which we observe non-zero trade flows in both directions, where every country is observed equally often in the sample, and where we observe trade flows

[8]The calculations were done on an x-64 based PC with Microsoft Windows 10 Pro operating system, the processor is Intel(R) Core(TM) i7-6920HQ CPU @2.90GHz, with 4 cores and 8 logical processors, 64 GB physical RAM.

[9]We specify the *maximum* number of fixed effects, acknowledging that some of the fixed effects are omitted from linear regressions due to collinearity. Note that ANOVA commands, such as `aov` in R, check collinearity and drop perfectly multicollinear variables, which takes time, included in the provided calculation times.

[10]The `anova` function in STATA tends to perform similarly to the `aov` function in R, generally being slightly slower.

[11]More details on the subsamples and further results can be found in Appendix C.

Table 1: Computing Time Improvement for Different (Sub-)Samples.

Sample	**Time** `aov`	**Time ANOVA-HDFE**	**Ratio** $\frac{\text{time aov}}{\text{time ANOVA-HDFE}}$
Strongly balanced cross-section (1998), 2,809 obs., 53 countries, implying at most 4,346 total FE	1 min 6 sec	0.12 sec	550
Weakly balanced cross-section (1998), 7,765 obs., 101 countries, implying at most 11,900 total FE	38 min 39 sec	0.13 sec	17,838
Strongly balanced 3-year panel (2011-13), 8,427 obs., 53 countries, implying at most 17,387 total FE	4 h 53 min 17 sec	0.14 sec	125,693
Weakly balanced 3-year panel (2011-13), 20,568 obs., 90 countries, implying at most 42,039 total FE	2 d 20 h 27 min 0 sec	0.16 sec	1,540,125
Strongly balanced 37-year panel (1980-2016), 7,252 obs., 14 countries, implying at most 12,539 total FE	53 min 42 sec	0.13 sec	24,785
Weakly balanced 37-year panel (1980-2016), 39,183 obs., 37 countries, implying at most 63,915 total FE	5 d 12 h 10 min 27 sec	0.22 sec	2,162,850
Entire unbalanced panel (1980-2016), 665,542 obs., 223 countries, implying at most 1,137,637 total FE	error: cannot allocate vector of size 1,172.4 GB	7.38 sec	

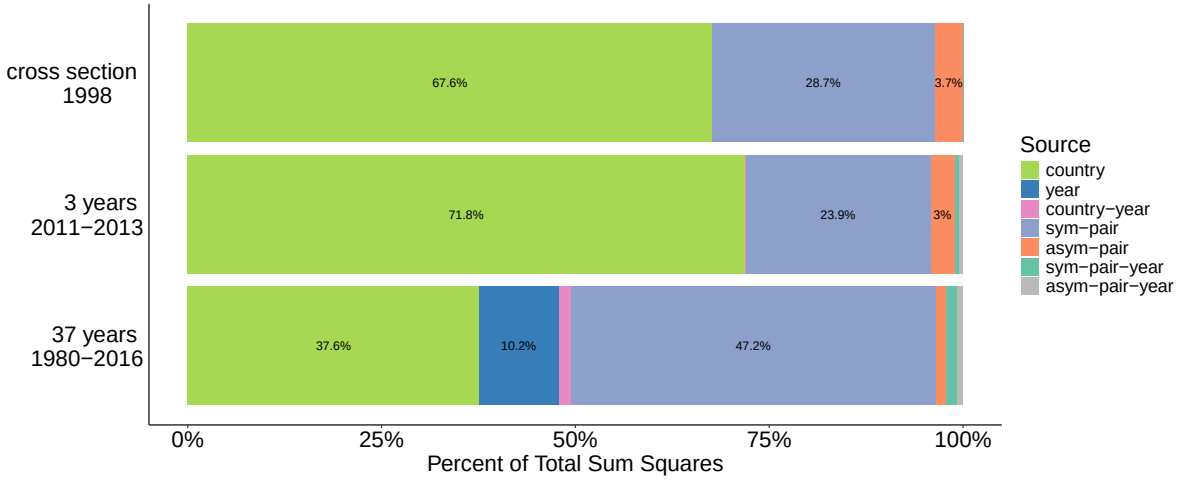
between all countries in all years of the respective sample.

As a starting point, we partition the variance of a cross-section of trade flows. The upper bar of Panel (a) in Figure 1 shows the results of an ANOVA decomposing the variation, measured as sums of squares, of the natural logarithm of trade flows for a strongly balanced cross-section in the year 1998 (the mid-year of the entire panel) with 53 countries. 67.6% of the variation in log trade flows is country-specific (i.e. explained by the exporter and importer fixed effects, which we group together, since there is no theoretical ordering on which set to include first), while 32.4% is pair-specific. Among this pair-specific variation, 89% is symmetric and 11% is asymmetric.[12] These results show that pair-specific variables play a substantial role in explaining trade flows in a cross-section, with a non-negligible role played by asymmetries. While in a cross-section, the time-invariant country and pair fixed effects explain the total variation in log trade flows, moving to a panel adds a time dimension. The middle bar of Panel (a) in Figure 1 shows results for a strongly balanced and short panel for 53 countries over the years 2011–2013. Over the three years, the year and country-year variation, as well as the bilateral time variation, are not substantial. Moving one step further, we decompose the variation for a strongly balanced subsample of the total 37-year period covered by our data (1980–2016), including 14 countries, in the lower bar of Panel (a) in Figure 1. 10.2% of the variation is year-specific, with the country-year fixed effects adding only 1.6 percentage points of explained variation to the explanatory power of the exporter, importer, and year fixed effects. The bilateral variation accounts for roughly half of the variation, of which the majority is symmetric and time-invariant.

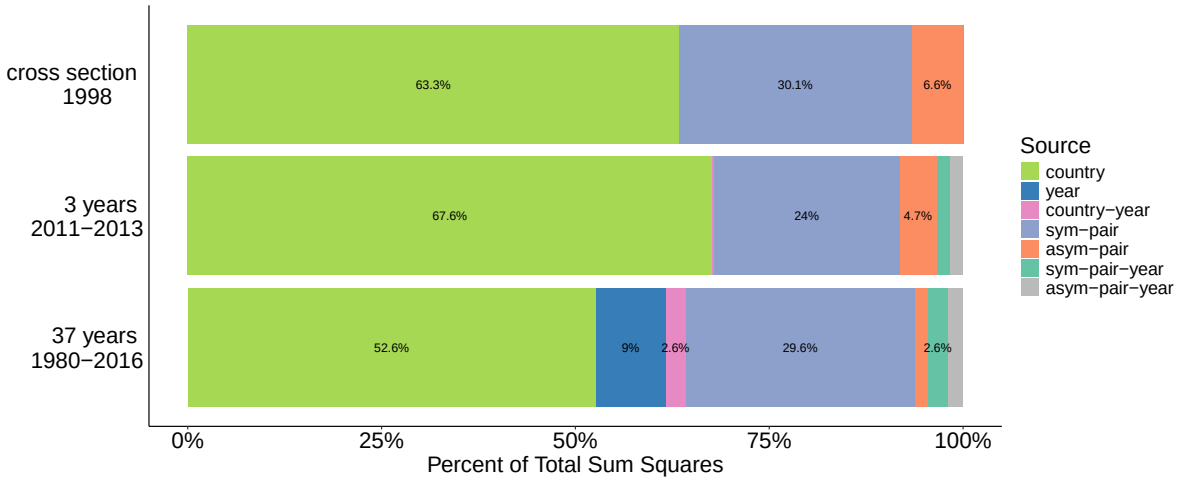
Strongly balanced subsamples rely on a specific set of countries with good availability of trade data, including mostly high-income countries, which are similar to each other in many respects. Hence, the sample is not particularly representative of worldwide trade flows, and certain components of variation may play a more important role than in a more representative worldwide sample. Therefore, in the next step, we move from strongly balanced subsamples to weakly balanced subsamples. In weakly balanced panels,

[12] All share calculations are given in Appendix D.

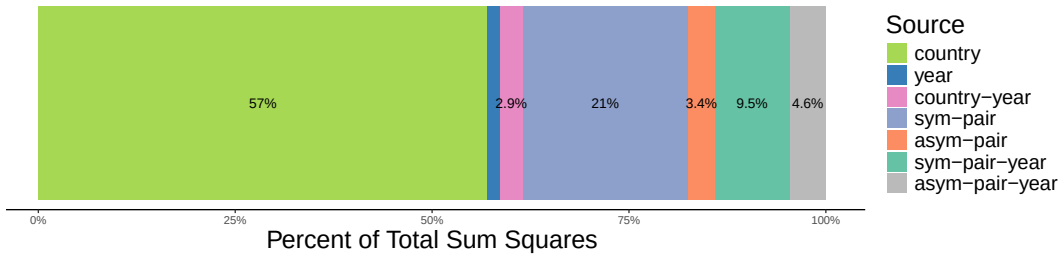
Figure 1: ANOVA of the natural logarithm of trade flows, different (sub-)samples, fixed effects capture all dimensions of variation.



(a) Strongly balanced subsamples.



(b) Weakly balanced subsamples.



(c) Entire panel dataset, years 1980–2016.

Notes: In Panel (a), the upper bar shows ANOVA results for a strongly balanced cross-section in 1998, the middle bar shows ANOVA results for a strongly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a strongly balanced 37-year panel. In Panel (b), the upper bar shows ANOVA results for a weakly balanced cross-section in 1998, the middle bar shows ANOVA results for a weakly balanced 3-year panel, 2011–2013, and the lower bar shows ANOVA results for a weakly balanced 37-year panel. Panel (c) shows ANOVA results for the entire unbalanced dataset, 1980–2016. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

we observe non-zero trade flows in both directions in every year for a country pair, but we do not observe all countries equally often, as we do not observe every country trading with every other country. Panel (b) in Figure 1 shows a similar pattern for the weakly balanced subsamples as for the strongly balanced subsamples, while asymmetries tend to play a larger role. The upper bar shows that the overall variation explained by the asymmetric bilateral component increases to 6.6% in the cross-section, while the middle bar shows that the time-invariant and time-varying asymmetric bilateral components together explain 6.4% in the weakly balanced 3-year panel. In the 37-year weakly balanced panel with 37 countries, the country variation increases while the time-invariant symmetric pair variation decreases compared to the strongly balanced panel, as can be seen in the lower bar of Panel (b) in Figure 1.

Overall, in all balanced subsamples, the country(-year)-specific component explains roughly two-thirds of the variation in all balanced subsamples, leaving one-third bilateral(-time) variation. In the cross-section, asymmetric pair fixed effects completely capture all the bilateral variation, whereas in the panel, there is some time-varying bilateral variation to be explained. We observe a non-negligible role played by bilateral asymmetries, especially time-invariant asymmetries, but there are also some time-varying asymmetries. Asymmetries appear to be more pronounced in the weakly balanced samples than in the strongly balanced samples, potentially explained by increased heterogeneity between the countries in weakly balanced panels and therefore less similarity to be captured by symmetric components.[13]

In practice, researchers often use all available trade data, which is unbalanced due to both missing reported trade data and zero trade flows, which are dropped when taking the natural logarithm. Panel (c) in Figure 1 shows the results of an ANOVA of the natural logarithm of trade flows for our entire unbalanced panel in the years 1980–2016 with

[13]The larger role played by asymmetries in the weakly balanced samples compared with the strongly balanced samples is not driven by the technical fact that the sample is weakly balanced (i.e. we are not introducing asymmetries technically by the departure from strong balancedness), but rather from the fact that we are adding more different countries (i.e. different income levels, etc.). We test this by starting with our strongly balanced samples and randomly dropping observations to make the samples weakly unbalanced. Running ANOVAs for these randomly created subsamples, we see that the asymmetries play a similar role as in the underlying strongly balanced samples.

223 countries. 57% of the variation is country-specific, while the time-specific component accounts for only 0.9% of the variation. We see that asymmetries matter, with 3.4% of the total variation being time-invariant asymmetric-pair specific and 4.6% being time-varying asymmetric-pair specific. Expressed as a share of the bilateral variation, about 21% are asymmetric.

5 ANOVA-HDFE With Covariates: Investigating the Explanatory Power of Trade Cost Proxies for Bilateral Trade Flows

In Section 4, we decomposed trade flows into their three dimensions—exporter, importer, and time—and their combinations, including bilateral (time) variation. The standard estimating gravity equation for trade flows is derived from theory and includes size and trade cost terms (see Head and Mayer, 2014; Yotov et al., 2016, for surveys). Having no sound theoretical foundation, the bilateral component of trade costs is typically proxied by observable variables. The standard gravity trade cost variables are the geographic distance between two countries, as well as indicator variables whether the countries share a contiguous border, whether they share a common language, or whether they have a common colonial heritage. Further trade cost proxy variables have been suggested in the literature, often referred to as policy variables (see Larch and Yotov, 2024). These include indicator variables for an RTA between the countries, for joint WTO membership, joint EU membership, and international borders (which capture the difference between international versus domestic trade). In a panel, the estimating equation can include symmetric pair fixed effects (i.e. the same fixed effect for flows between a pair in both directions) or asymmetric pair fixed effects (i.e. an own fixed effect for each direction of flows between a pair), which capture all bilateral trade cost proxy variables that are constant over time, especially the gravity variables, such as geographic distance, leaving only time-varying bilateral variables in the trade costs.

In gravity equations, any variation in trade flows beyond the size and price terms

(which are captured by exporter-time and importer-time fixed effects) is attributed to bilateral trade costs. Hence, trade costs are a main explanatory factor of trade flows (see, for example, Chaney, 2018), albeit without being directly observable. To assess the explanatory power of standard trade cost proxies, we decompose the bilateral variation of the natural logarithm of trade flows into the different components using our ANOVA-HDFE procedure. For comparison reasons, we first show the different dimensions of variation using fixed effects as in Section 4, focusing solely on the bilateral component of variation, before including trade cost proxy variables in a later step. Panel (a) in Figure 2 shows the components of variation in the bilateral variation in the logarithm of trade flows for weakly balanced subsamples.[14] Most of the bilateral variation (between 75 and 83%) is symmetric time-invariant. Asymmetries (both constant and varying over time) account for about 10 to 20% of the bilateral variation. Panel (a) in Figure 3 shows the components for the entire unbalanced panel, where about 55% of the variation is symmetric and time-invariant, while 21% is asymmetric (both constant and varying over time). In the full panel, 37% of the bilateral variation varies over time.

Next, we add the trade cost proxy variables to the ANOVA calculation, where we first add the standard gravity variables (distance, contiguity, common language, and common colonial heritage), before adding the policy variables (RTAs, joint WTO membership, joint EU membership, and international borders), and then the fixed effects. As outlined in Section 2, the ordering of the variables added to the ANOVA calculation using sequential sums of squares matters. The common variation of the variables with the dependent variable is assigned to the variable that is included first. However, a group of variables together always explains the same amount of variation, no matter how the variables are ordered within the group. As there is no theory on the relative importance of the trade cost proxy variables, we report the results in groups, first for the time-invariant gravity variables and then for the policy variables, given that the latter have some time variation going beyond the variation of the gravity variables.

Panel (b) of Figure 2 shows that, in weakly balanced subsamples, among the symmetric

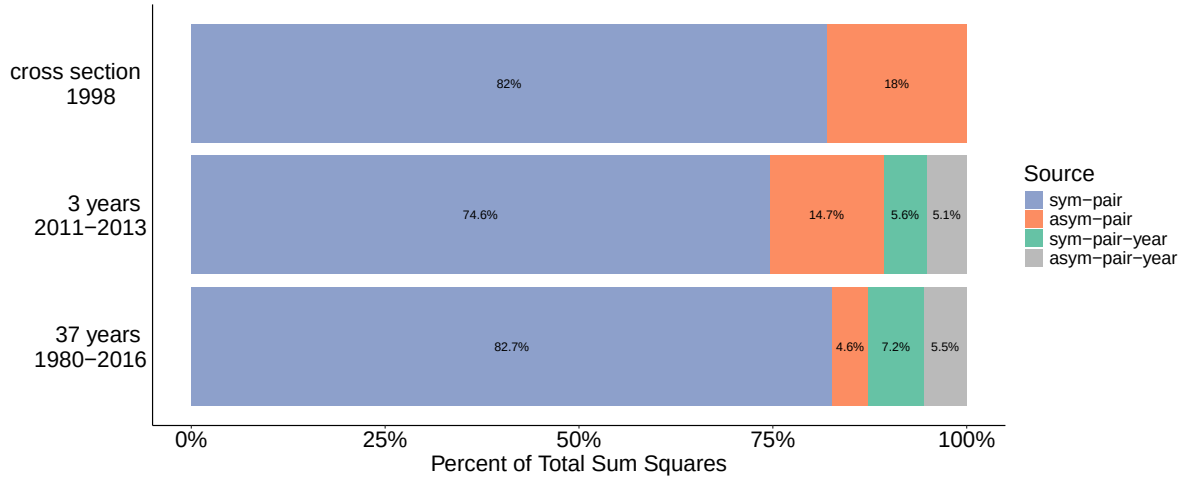
[14]Results for strongly balanced subsamples deliver a similar pattern and are shown in Figure C.3 in the Appendix.

time-invariant variation—which is essentially the dimension in which all the standard trade cost proxies vary—the standard gravity variables explain 58 to 74%. Adding the policy variables increases the overall explanatory power by about 2 to 6 percentage points. From Panel (a) to Panel (b) of Figure 2, the shares of variation captured by the asymmetric pair fixed effects do not change, as by construction, neither the gravity nor the policy variables have asymmetric variation (e.g. the geographic distance between two countries is the same in both directions, or if one country has an RTA with another, this trade agreement holds for trade in both directions). Varying over time, the policy variables do, however, capture some of the symmetric time-varying bilateral variation—this share decreases for the 37-year subsample from 7.2 to 6.8%. In the entire unbalanced panel dataset in Panel (b) of Figure 3, 56% of the symmetric time-invariant variation is explained by the standard gravity variables. The policy variables increase the overall explanatory power by about 1 percentage point, among which there is only little time variation, since the symmetric time variation decreases by only 0.1 percentage points from 24.7% in Panel (a) to 24.6% in Panel (b).

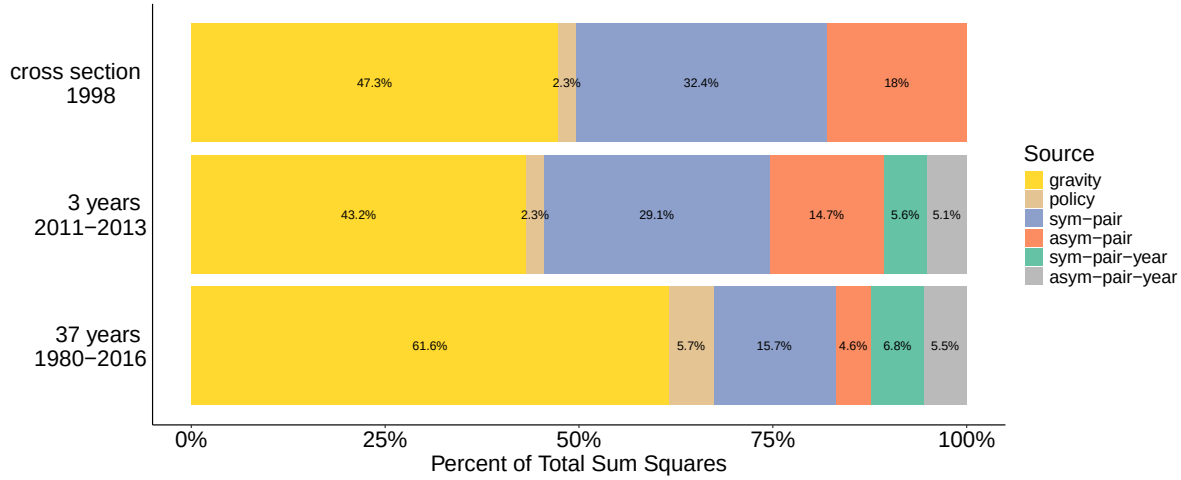
Overall, although the trade cost proxy variables together do explain a substantial part of the variation in bilateral trade costs (32 to 67%), they also leave a substantial part unexplained. Of this remaining variation, 33 to 64% is symmetric and time-invariant, showing that the trade cost proxy variables not only fail to explain asymmetries by construction, but even the time-constant symmetric pair variation is not entirely captured. The substantial amount of asymmetries in trade costs shows that it may be fruitful to include asymmetric trade cost proxy variables, beyond the symmetric gravity and policy trade cost proxies.[15] From these results, we highlight three insights. First, after the inclusion of the trade cost proxy variables, a substantial part of time-invariant bilateral variation remains unexplained. Second, within the remaining unexplained variation, asymmetries matter. Third, considering that cross-sectional estimations cannot include pair fixed effects and hence they rely on trade cost proxy variables, they leave a substantial amount

[15] A straightforward asymmetric trade policy variable is tariffs, which can be imposed by one country on imports from another, while the other country may impose a different tariff rate. Partly due to data availability and data handling, tariffs are not yet included as a standard trade cost proxy variable in the gravity literature.

Figure 2: ANOVA of the bilateral variation in the natural logarithm of trade flows for weakly balanced subsamples.



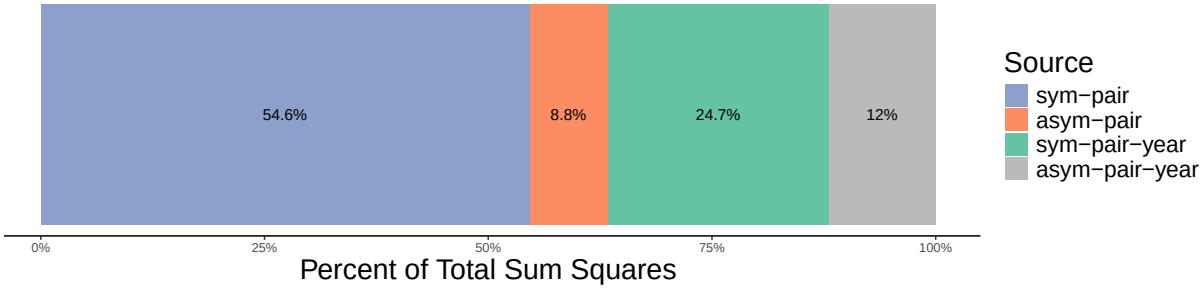
(a) ANOVA of the bilateral variation in the natural logarithm of trade flows, weakly balanced subsamples.



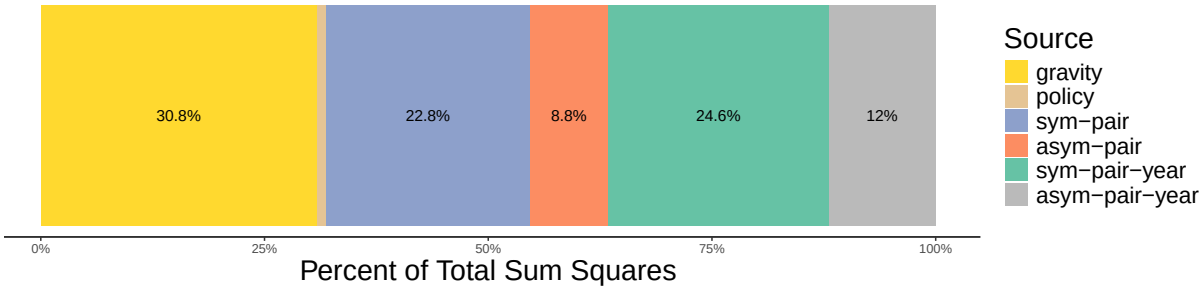
(b) ANOVA of the bilateral variation in the natural logarithm of trade flows, weakly balanced subsamples, with trade cost proxy variables.

Notes: In both panels, the upper bar shows ANOVA results for a balanced cross-section in 1998, the middle bar shows ANOVA results for a balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

Figure 3: ANOVA of the bilateral variation in the natural logarithm of trade flows, for the entire unbalanced dataset, years 1980–2016.



(a) ANOVA of the bilateral variation in the natural logarithm of trade flows.



(b) ANOVA of the bilateral variation in the natural logarithm of trade flows, with trade cost proxy variables.

Notes: Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

of trade costs (e.g. 50% in the weakly balanced cross-section in Panel (b) of Figure 2) unexplained and hence in the residual of gravity estimations. This residual can be substantially reduced by the use of bilateral pair fixed effects (e.g. to less than 25% in the weakly balanced panels in Panel (a) of Figure 2).

6 Conclusion

With ANOVA-HDFE, we propose a new procedure for conducting analyses of variance in the presence of high-dimensional fixed effects and provide a specific application to the analysis of trade flows. We contribute to both the literature on analysis of variance and the more specific literature analyzing the determinants of bilateral trade flows. Our procedure extends existing approaches for computing ANOVAs to accommodate the presence of high-dimensional fixed effects, which are increasingly prevalent in various settings, including trade, labor market, or firm-level analyses. ANOVA-HDFE combines sequential linear regressions and recent developments in regression analysis with high-dimensional fixed effects, making it substantially faster than existing procedures in decomposing the variation of a variable with many dimensions. For very large datasets or very many variables, ANOVA-HDFE offers a feasible procedure, where existing procedures fail to deliver results. Being fast and applicable with large data and high-dimensional fixed effects, ANOVA-HDFE is a recommendable tool for practitioners to obtain an impression of the variation in their data.

In terms of our application, we extend existing investigations of determinants of trade, using a broader dataset, and provide an overview of the nature of bilateral trade costs. Conducting an ANOVA on trade flows shows that about 60% of the variation in trade volumes is at the unilateral country(-time) level, a further 30% across different country pairs, and the remainder within different country pairs. Our analysis reveals that standard trade cost proxy variables, such as bilateral distance or trade agreements, leave a substantial proportion of pair-specific variation in international trade flows unexplained, and also that, within country pairs, there are large asymmetries in international trade costs, above and beyond the symmetric barriers to trade. We leave a more in-depth anal-

ysis of trade costs and the exploration of the potential sources of the asymmetries which we uncovered using ANOVA-HDFE for future research.

References

Bergé, L. (2018): "Efficient Estimation of Maximum Likelihood Models with Multiple Fixed-Effects: The R Package FENmlm," *CREA Discussion Paper 2018-13*.

Borchert, I., M. Larch, S. Shikher, and Y. Yotov (2022): "Disaggregated Gravity: Benchmark Estimates and Stylized Facts From a New Database," *Review of International Economics*, 30, 113–136.

Chambers, J. M. and T. J. Hastie (1992): *Statistical Models in S*, London, UK: Chapman & Hall.

Chaney, T. (2018): "The Gravity Equation in International Trade: An Explanation," *Journal of Political Economy*, 126, 150–177.

Cornelissen, T. (2018): "The Stata Command Felsdvreg to Fit a Linear Model with Two High-Dimensional Fixed Effects," *The Stata Journal*, 8, 170–189.

Correia, S. (2016): "A Feasible Estimator for Linear Models with Multi-Way Fixed Effects," *Unpublished Manuscript, Duke University*.

Cuñat, A. and R. Zymek (2024): "Bilateral Trade Imbalances," *Review of Economic Studies*, 91, 1537–1583.

Egger, P. H. and S. Nigai (2015): "Structural Gravity With Dummies Only: Constrained ANOVA-type Estimation of Gravity Models," *Journal of International Economics*, 97, 86–99.

Fisher, R. A. (1925): *Statistical Methods for Research Workers*, Edinburgh: Oliver & Boyd.

Gaure, S. (2013a): "lfe: Linear Group Fixed Effects," *R Journal*, 5, 104–116.

——— (2013b): "OLS with Multiple High Dimensional Category Variables," *Computational Statistics & Data Analysis*, 66, 8–18.

Gervais, A. (2025): "A Decomposition of the Variance of International Trade Flows," *Empirical Economics*, 68, 2073–2092.

Guimarães, P. and P. Portugal (2010): "A Simple Feasible Procedure to Fit Models With High-Dimensional Fixed Effects," *Stata Journal*, 10, 628–649.

Head, K. and T. Mayer (2014): "Gravity Equations: Workhorse, Toolkit, and Cookbook," in *Handbook of International Economics*, ed. by G. Gopinath, E. Helpman, and K. Rogoff, Elsevier B.V., vol. 4, 131–195.

Kendall, M., A. Stuart, and J. K. Ord (1983): *The Advanced Theory of Statistics, Volume 3: Design and Analysis, and Time Series*, London & High Wycombe: Charles Griffin & Company Limited, 4 ed.

Larch, M., J. Monteiro, R. Piermartini, and Y. V. Yotov (2025): "On the Trade Effects of GATT/WTO Membership: They are Positive and Large After All," *Canadian Journal of Economics/Revue canadienne d'économique*, 58, 281–328.

Larch, M., J. Wanner, Y. Yotov, and T. Zyklin (2019): "Currency Unions and Trade: A PPML Re-Assessment with High-Dimensional Fixed Effects," *Oxford Bulletin of Economics and Statistics*, 81, 487–510.

Larch, M. and Y. V. Yotov (2024): "Estimating the Effects of Trade Agreements: Lessons from 60 Years of Methods and Data," *World Economy*, 47, 1771–1799.

Mittag, K. (2019): "A Simple Method to Estimate Large Fixed Effects Models Applied to Wage Determinants," *Labour Economics*, 61, 101766.

Redding, S. J. and D. E. Weinstein (2024): "Accounting for Trade Patterns," *Journal of International Economics*, 150, 103910.

Rommel, T. and P. Schaudt (2020): "First Impressions: How Leader Changes Affect Bilateral Aid," *Journal of Public Economics*, 185, 104107.

Sahai, H. and M. I. Ageel (2000): *The Analysis of Variance: Fixed, Random and Mixed Models*, Boston, MA: Birkhäuser.

Scheffé, H. (1999): *The Analysis of Variance*, Wiley.

Searle, S. R., G. Casella, and C. E. McCulloch (2006): *Variance Components*, Wiley.

Torres, S., P. Portugal, J. Addison, and P. Guimaraes (2018): "The Sources of Wage Variation and the Direction of Assortative Matching: Evidence from a Three-Way High-Dimensional Fixed Effects Regression Model," *Labour Economics*, 54, 47–60.

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

Zhang, D., D. Bai, and Y. Wang (2024): "Green vs. Brown: Climate Risk Showdown–Who's Thriving, Who's Diving?" *Journal of International Money and Finance*, 149, 103198.

Zhu, X., X. Zuo, and H. Li (2021): "The Dual Effects of Heterogeneous Environmental Regulation on the Technological Innovation of Chinese Steel Enterprises—Based on a High-Dimensional Fixed Effects Model," *Ecological Economics*, 188, 107113.

Online Appendix

A Details on ANOVA Theory

The total sum of squares of the three-way variable of interest, SQ_{total}, is the sum of squared deviations of the variable's observations from the grand mean $\overline{X}_{...}$, that is, from the mean over all observations of X_{ijt}. It can be partitioned into the following components:

$$SQ_{total} = SQ_i + SQ_j + SQ_t + SQ_{it} + SQ_{jt} + SQ_{\underset{ij}{\leftrightarrow}} + SQ_{ij} + SQ_{\underset{ij}{\leftrightarrow}t} + SQ_{ijt},$$

with

$$\begin{aligned}
SQ_{total} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\left(X_{ijt} - \overline{X}_{...}\right)^2, \\
SQ_i &= JT\sum_{i=1}^{I}\left(\overline{X}_{i..} - \overline{X}_{...}\right)^2, \\
SQ_j &= IT\sum_{j=1}^{J}\left(\overline{X}_{.j.} - \overline{X}_{...}\right)^2, \\
SQ_t &= IJ\sum_{t=1}^{T}\left(\overline{X}_{..t} - \overline{X}_{...}\right)^2, \\
SQ_{it} &= J\sum_{i=1}^{I}\sum_{t=1}^{T}\left(\overline{X}_{i\cdot t} - \overline{X}_{...} - (\overline{X}_{i..} - \overline{X}_{...}) - \left(\overline{X}_{..t} - \overline{X}_{...}\right)\right)^2 \\
&= J\sum_{i=1}^{I}\sum_{t=1}^{T}\left(\overline{X}_{i\cdot t} - \overline{X}_{i..} - \overline{X}_{..t} + \overline{X}_{...}\right)^2, \\
SQ_{jt} &= I\sum_{j=1}^{J}\sum_{t=1}^{T}\left(\overline{X}_{.jt} - \overline{X}_{...} - (\overline{X}_{.j.} - \overline{X}_{...}) - \left(\overline{X}_{..t} - \overline{X}_{...}\right)\right)^2 \\
&= I\sum_{j=1}^{J}\sum_{t=1}^{T}\left(\overline{X}_{.jt} - \overline{X}_{.j.} - \overline{X}_{..t} + \overline{X}_{...}\right)^2, \\
SQ_{\underset{ij}{\leftrightarrow}} &= T\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{\underset{ij}{\leftrightarrow}.} - \overline{X}_{...} - \left(\overline{\overline{X}_{i..}}_{\underset{ij}{\leftrightarrow}.} - \overline{X}_{...}\right) - \left(\overline{\overline{X}_{.j.}}_{\underset{ij}{\leftrightarrow}.} - \overline{X}_{...}\right)\right)^2 \\
&= T\sum_{i=1}^{I}\sum_{j=1}^{J}\left(\overline{X}_{\underset{ij}{\leftrightarrow}.} - \overline{\overline{X}_{i..}}_{\underset{ij}{\leftrightarrow}.} - \overline{\overline{X}_{.j.}}_{\underset{ij}{\leftrightarrow}.} + \overline{X}_{...}\right)^2,
\end{aligned}$$

$$
\begin{aligned}
SQ_{ij} &= T\sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\overline{X}_{ij\cdot}-\overline{X}_{\cdots}-\big(\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdots}\big)-\big(\overline{X}_{\cdot j\cdot}-\overline{X}_{\cdots}\big)\\
&\quad-\Big(\overline{X}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{X}_{\cdots}\Big)\Big)^2\\
&= T\sum_{i=1}^{I}\sum_{j=1}^{J}\Big(\overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}-\overline{X}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\Big)^2,\\
SQ_{\overleftrightarrow{ij}t} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\Big(\overline{X}_{\overleftrightarrow{ij}t}-\overline{X}_{\cdots}-\Big(\overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t}-\overline{X}_{\cdots}\Big)-\Big(\overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t}-\overline{X}_{\cdots}\Big)\\
&\quad+\big(\overline{X}_{\cdot\cdot t}-\overline{X}_{\cdots}\big)-\Big(\overline{X}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{X}_{\cdots}\Big)\Big)^2\\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\Big(\overline{X}_{\overleftrightarrow{ij}t}-\overline{X}_{\cdots}-\overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t}-\overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t}+\overline{X}_{\cdot\cdot t}-\overline{X}_{\overleftrightarrow{ij}\cdot}\\
&\quad+\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\Big)^2,\\
SQ_{ijt} &= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\Big(X_{ijt}-\overline{X}_{\cdots}-\big(\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdots}\big)-\big(\overline{X}_{\cdot j\cdot}-\overline{X}_{\cdots}\big)\\
&\quad-\big(\overline{X}_{\cdot\cdot t}-\overline{X}_{\cdots}\big)\\
&\quad-\big(\overline{X}_{i\cdot t}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot\cdot t}+\overline{X}_{\cdots}\big)-\big(\overline{X}_{\cdot jt}-\overline{X}_{\cdot j\cdot}-\overline{X}_{\cdot\cdot t}+\overline{X}_{\cdots}\big)\\
&\quad-\Big(\overline{X}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{X}_{\cdots}\Big)\\
&\quad-\Big(\overline{X}_{ij\cdot}-\overline{X}_{i\cdot\cdot}-\overline{X}_{\cdot j\cdot}-\overline{X}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\Big)\\
&\quad-\Big(\overline{X}_{\overleftrightarrow{ij}t}-\overline{X}_{\cdots}-\overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t}-\overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t}+\overline{X}_{\cdot\cdot t}-\overline{X}_{\overleftrightarrow{ij}\cdot}\\
&\quad+\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}+\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}\Big)\Big)^2\\
&= \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}\Big(X_{ijt}+\overline{X}_{i\cdot\cdot}+\overline{X}_{\cdot j\cdot}-\overline{X}_{i\cdot t}-\overline{X}_{\cdot jt}+\overline{X}_{\overleftrightarrow{ij}\cdot}\\
&\quad-\overline{\overline{X}_{i\cdot\cdot}}_{\overleftrightarrow{ij}\cdot}-\overline{\overline{X}_{\cdot j\cdot}}_{\overleftrightarrow{ij}\cdot}-\overline{X}_{ij\cdot}-\overline{X}_{\overleftrightarrow{ij}t}+\overline{\overline{X}_{i\cdot t}}_{\overleftrightarrow{ij}t}+\overline{\overline{X}_{\cdot jt}}_{\overleftrightarrow{ij}t}\Big)^2,
\end{aligned}
$$

where

$$
\begin{aligned}
\overline{X}_{\cdots} &= \frac{1}{IJT}\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{t=1}^{T}X_{ijt},\\
\overline{X}_{i\cdot\cdot} &= \frac{1}{JT}\sum_{j=1}^{J}\sum_{t=1}^{T}X_{ijt},
\end{aligned}
$$

$$
\begin{aligned}
\overline{X}_{\cdot j \cdot} &= \frac{1}{IT}\sum_{i=1}^{I}\sum_{t=1}^{T} X_{ijt}, \\
\overline{X}_{\cdot\cdot t} &= \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} X_{ijt}, \\
\overline{X}_{i \cdot t} &= \frac{1}{J}\sum_{j=1}^{J} X_{ijt}, \\
\overline{X}_{\cdot jt} &= \frac{1}{I}\sum_{i=1}^{I} X_{ijt}, \\
\overline{X}_{\overleftrightarrow{ij} \cdot} &= \begin{cases} \frac{1}{T}\sum_{t=1}^{T} X_{ijt}, & \text{if } i = j \\ \frac{1}{2T}\sum_{t=1}^{T}\sum_{i,j\in p} X_{ijt}, & \text{otherwise}, \forall p \in P \end{cases}, \\
\overline{X}_{ij \cdot} &= \frac{1}{T}\sum_{t=1}^{T} X_{ijt}, \\
\overline{X}_{\overleftrightarrow{ij}\, t} &= \begin{cases} X_{ijt}, & \text{if } i = j \\ \frac{1}{2}\sum_{i,j\in p} X_{ijt}, & \text{otherwise}, \forall p \in P \end{cases},
\end{aligned}
$$

where p are the pairs in the set of pairs P. That is, for the symmetric-pair(-time), we sum over the directional observations for each symmetric pair. Variables with two bars denote means of means, e.g. $\overline{\overline{X}_{i\cdot\cdot\overleftrightarrow{ij}\cdot}}$ are the means over $\overleftrightarrow{ij} \cdot \left(\overline{X}_{\overleftrightarrow{ij}\cdot}\right)$ of the means over $i \cdot \cdot \left(\overline{X}_{i\cdot\cdot}\right)$.

B ANOVA Implementation in R

In this section we describe how ANOVAs can be implemented, using the method implemented in the program R and closely following the exposition in Chambers and Hastie, 1992, Chapters 4 and 5.

Essentially, the sums of squares are computed as the squared *effects*. In the following, we explain what the effects are and how they are used to calculate sequential sums of squares in an ANOVA.[16] The effects result from a QR decomposition that is routinely undertaken when a linear model is estimated in R. This method of calculating the sums of squares relying on QR decomposition is also what operates in the background of the `aov` function used for fitting ANOVAs in R.

In the orthogonal transformation underlying the QR decomposition, the $(N \times K)$ matrix of regressors $\mathbf{X}$ is decomposed as

$$\mathbf{X} = \mathbf{QR},$$

where $\mathbf{R}$ is a $(K \times K)$ upper triangular matrix and $\mathbf{Q}$ is an $(N \times K)$ matrix with rank K. That is, the columns of $\mathbf{Q}$ are linearly independent and

$$\mathbf{Q}' \cdot \mathbf{Q} = \mathbf{I},$$

where $\mathbf{Q}'$ is the transpose of $\mathbf{Q}$ and $\mathbf{I}$ is the $(K \times K)$ identity matrix.

Geometrically, $\mathbf{Q}$ is a set of orthogonal vectors that span the columns of $\mathbf{X}$. Any linear combination of the columns of $\mathbf{X}$ can be written as a linear combination of the columns of $\mathbf{Q}$.

The effects are given by the vector $\mathbf{c}$ of length K satisfying

$$\mathbf{c} = \mathbf{Q}' \cdot \mathbf{y},$$

[16] We give a brief and practical outline, in line with the commands in R, while further details can be found in Chambers and Hastie, 1992, Chapter 4.

where $\mathbf{y}$ is the vector of the dependent variable of length N.
The sum of squares explained by the individual regressors is obtained by squaring the effects in $\mathbf{c}$,

$$SQ = \mathbf{c}^2.$$

In summary, all one needs to calculate the sums of squares is the matrix $\mathbf{Q}$, resulting from QR decomposition, and the dependent variable vector $\mathbf{y}$. This can easily be implemented in R: Create a matrix of regressors `X`, including a constant and an indicator variable for each fixed effect, but excluding any reference-group indicator variable that would be omitted from a linear regression due to collinearity. Order the variables in the order in which they are to be included in the sequential sums of squares, starting with the intercept. It is advisable to add column names to address and interpret the variables in later steps. Create a vector of the dependent variable, `y` (e.g. consisting of log trade flows). Perform the QR decomposition and extract the matrix `Q`. Using `Q` and `y`, calculate the `effects` and square them to obtain the sums of squares `SQ`.

```
QR <- qr(X)
Q <- qr.Q(QR)
effects <- t(Q)%*%y
SQ <- effects^2
```

`SQ` gives the effects for all variables included in the matrix `X`. It does not include the residual sum of squares, which is usually reported by ANOVA outputs, while it includes the sum of squares explained by the intercept, which is usually not reported in the corrected total sum of squares in ANOVA outputs (Chambers and Hastie, 1992, p. 183). The residual sum of squares can easily be computed by subtracting the explained sum of squares (without the part assigned to the intercept) from the total sum of squares:

```
sum((y-mean(y))^2)-sum(SQ[2:length(SQ)])
```

In case the regressor of interest is a group of variables (e.g. the exporter fixed effects, consisting of an indicator variable for each exporter), the sum of squares is obtained by adding the individually squared effects obtained for each variable. The method using QR

decomposition in R works in both cases, the balanced as well as the unbalanced one, and yields the same results as the `aov` command in R, the `anova` command in STATA, and, in the balanced case, as our method from section 2.1.

In practice, the researcher does not necessarily have to perform a QR decomposition and can instead rely on R's command `lm`, used for estimating linear regressions, which routinely performs a QR decomposition in the background.[17] Since `lm` can handle factor variables, this has the advantage that the indicator variables for the fixed effects do not all have to be created as their own indicator variables. As such, the creation of the matrix of regressors $\mathbf{X}$ with all relevant dummies can be avoided. To do so, create factor variables for the fixed effects using the `as.factor` command in R. Use these factor variables in a linear regression `lm` and extract the `effects` from the obtained object, for example

```
data$exp <- as.factor(data$exporter)
data$imp <- as.factor(data$importer)
lmfactor <- lm(lntrade ~ exp + imp + common_language, data=data)
lmfactor$effects
```

The first elements in the obtained `effects` are the effects associated with the respective regressor (notice that an effect is given for each level of the factor variable fixed effects). As above, they can be squared to obtain the sum of squares and potentially added for the groups of fixed effects. The remaining elements in `effects` span the space of the residuals.

Unfortunately, though, for the computation of the effects the researcher cannot rely directly on fast-fixed-effects linear estimation methods implemented, for instance, in the package `lfe` by Gaure (2013a), since they do not include a QR-decomposition. Hence, even though fixed effects do not have to be created as their own indicator variables when using standard ANOVA commands such as `aov` or `lm` in R, each fixed effect has to be created by a standard computation method, instead of relying on faster implementations. Thus, ANOVAs for panel trade data with many dimensions, such as trade flows, including

[17] Here we explain how an ANOVA can be conducted without relying on pre-implemented functions in R that directly calculate ANOVAs. A straightforward option for the researcher could, obviously, also be to rely on the `aov` function used for fitting analysis of variance models in R. The QR decomposition is operated in the background.

exporter, importer, exporter-time, importer-time, symmetric pair, and asymmetric pair fixed effects, may be computationally demanding and entail long running times, without the possibility of parallelization.

C Details on Subsamples and Additional ANOVA Results

Strongly balanced subsamples: We observe every country trading with itself and with every other country every year. Details on the subsamples:

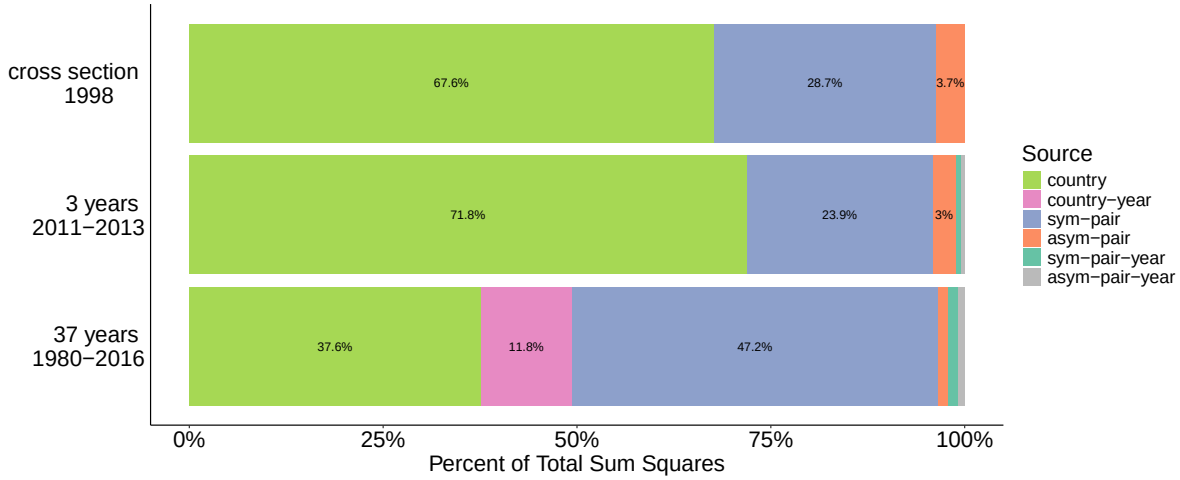
- strongly balanced cross-section 1998: 53 countries, 2809 observations, includes domestic trade,
- strongly balanced 3-year panel 2011-2013: 53 countries, 8427 observations in total, 2809 observations per year, includes domestic trade,
- strongly balanced 37-year panel 1980-2016: 14 countries, 7252 observations in total, 196 observations per year, includes domestic trade, countries: AUS, CAN, DNK, FRA, GRC, IRL, ITA, JPN, NLD, NOR, PRT, SGP, ESP, SWE.

Panel (a) of Figure 1 shows results for strongly balanced subsamples, where the fixed effects capture all dimensions of variation. Figure C.1 shows results for strongly balanced subsamples, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation. Figure C.2 shows results for strongly balanced subsamples, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation, with gravity and policy trade cost proxy variables. Panel (a) of Figure C.3 shows results for the bilateral variation in the logarithm of trade flows for strongly balanced subsamples. Panel (b) of Figure C.3 shows results for the bilateral variation in the logarithm of trade flows for strongly balanced subsamples, with gravity and policy variables.

Weakly balanced subsamples: We observe every trading pair in both directions every year, but we do not observe every country trading with every other country, i.e. the trade matrix has the same missing pair observations each year. Includes domestic trade. Details on the subsamples:

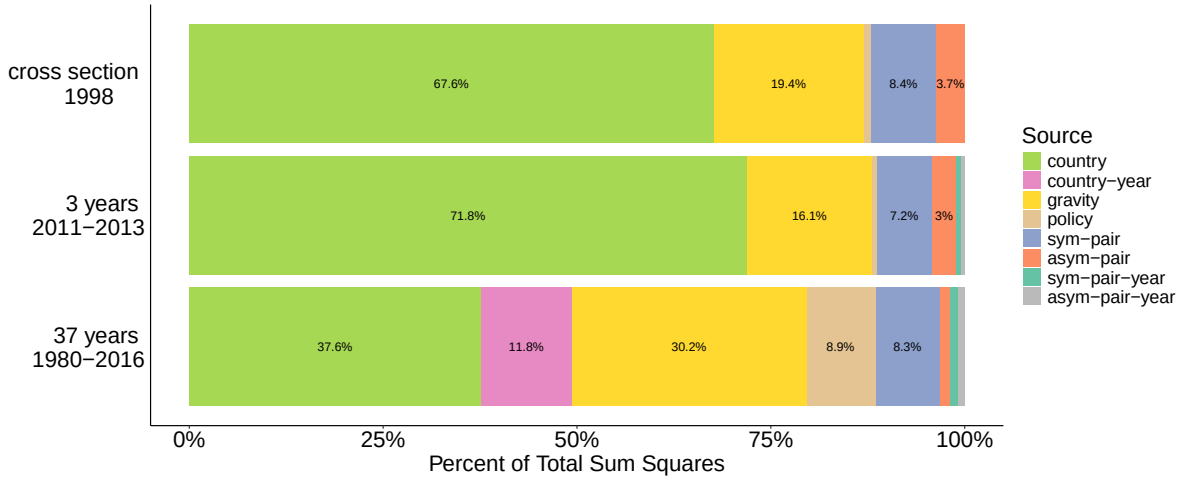
- weakly balanced cross-section 1998: 101 countries, 7765 observations, includes domestic trade,

Figure C.1: ANOVA of the natural logarithm of trade flows, strongly balanced subsamples, standard fixed effects to capture size and MRT terms.



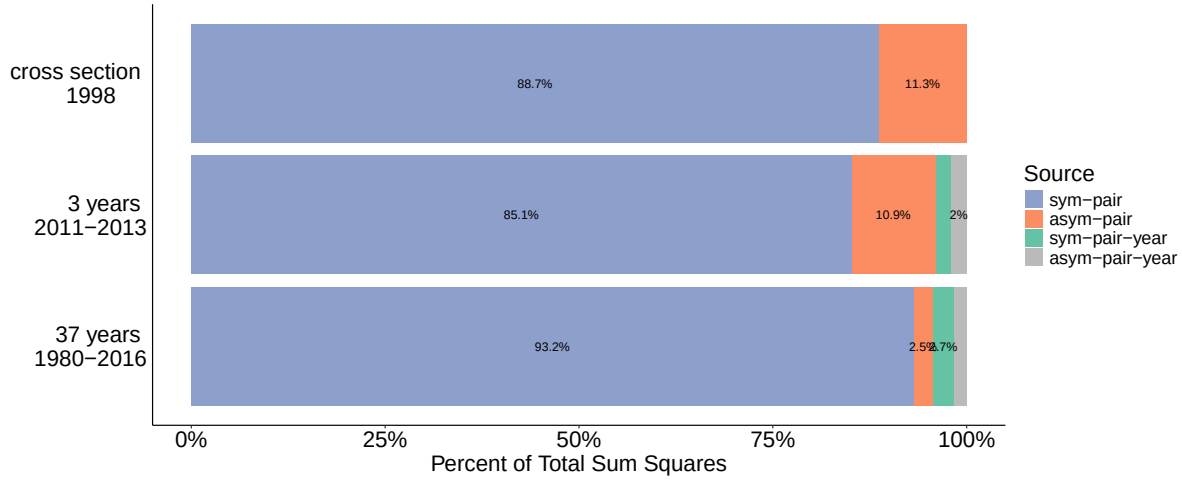
Notes: The upper bar shows ANOVA results for a strongly balanced cross-section in 1998, the middle bar shows ANOVA results for a strongly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a strongly balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

Figure C.2: ANOVA of the natural logarithm of trade flows, strongly balanced subsamples, standard fixed effects to capture size and MRT terms, with gravity trade cost proxy variables.

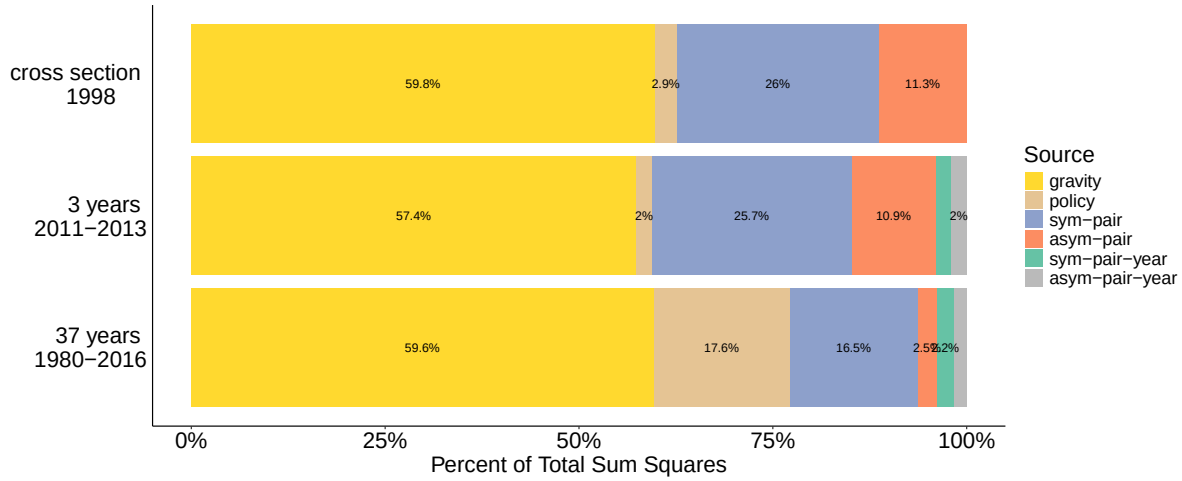


Notes: The upper bar shows ANOVA results for a strongly balanced cross-section in 1998, the middle bar shows ANOVA results for a strongly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a strongly balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

Figure C.3: ANOVA of the bilateral variation in the natural logarithm of trade flows for strongly balanced subsamples.



(a) ANOVA of the bilateral variation in the natural logarithm of trade flows, strongly balanced subsamples.



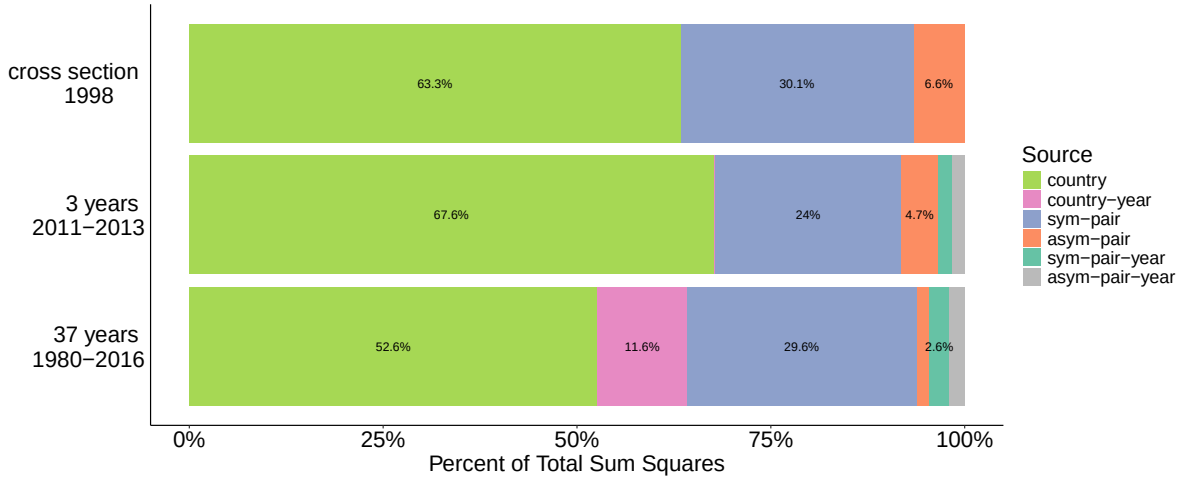
(b) ANOVA of the bilateral variation in the natural logarithm of trade flows, strongly balanced subsamples, with trade cost proxy variables.

Notes: In both panels, the upper bar shows ANOVA results for a strongly balanced cross-section in 1998, the middle bar shows ANOVA results for a strongly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a strongly balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

- weakly balanced 3-year panel 2011-2013: 90 countries, 6856 observations per year, 20568 observations in total, includes domestic trade,
- weakly balanced 37-year panel 1980-2016: 37 countries, 39183 observations in total, 1059 observations per year, includes domestic trade.

Panel (b) of Figure 1 shows results for weakly balanced subsamples, where the fixed effects capture all dimensions of variation. Figure C.4 shows results for weakly balanced subsamples, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation. Figure C.5 shows results for weakly balanced subsamples, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation, with gravity and policy trade cost proxy variables. Panel (a) of Figure 2 shows results for the bilateral variation in the logarithm of trade flows for weakly balanced subsamples. Panel (b) of Figure 2 shows results for the bilateral variation in the logarithm of trade flows for weakly balanced subsamples, with gravity and policy variables.

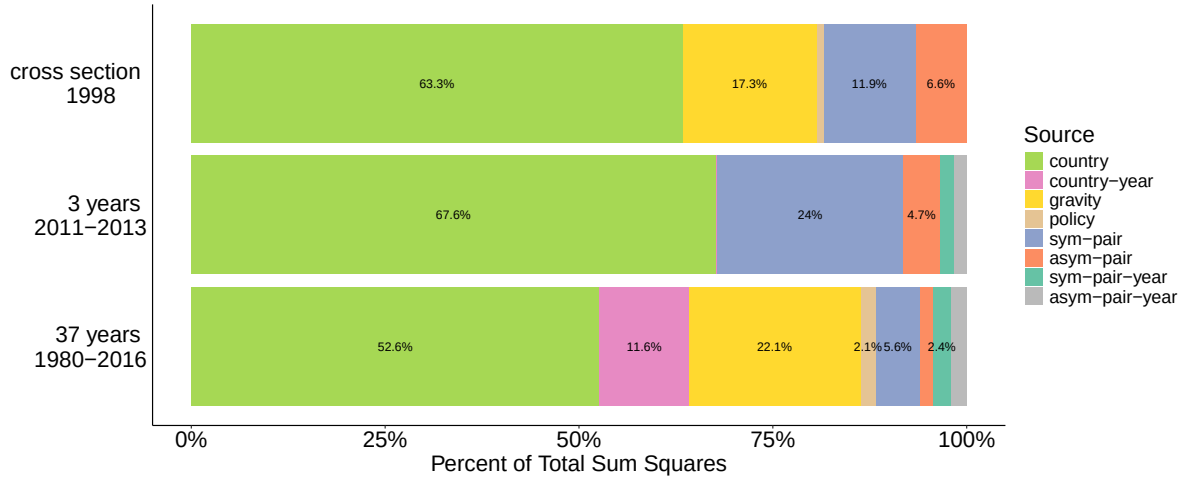
Figure C.4: ANOVA of the natural logarithm of trade flows, weakly balanced subsamples, standard fixed effects to capture size and MRT terms.



Notes: The upper bar shows ANOVA results for a weakly balanced cross-section in 1998, the middle bar shows ANOVA results for a weakly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a weakly balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

Entire panel dataset: 223 countries, 665542 observations in total, different number of observations per year, domestic trade not for every country and year.

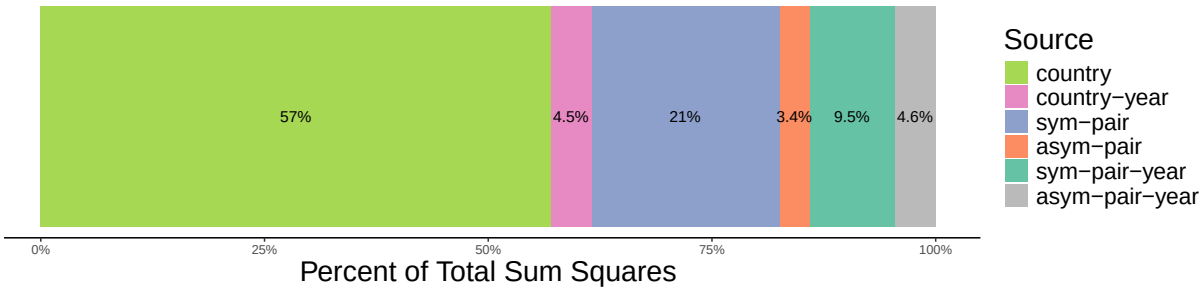
Figure C.5: ANOVA of the natural logarithm of trade flows, weakly balanced subsamples, standard fixed effects to capture size and MRT terms, with gravity trade cost proxy variables.



Notes: The upper bar shows ANOVA results for a weakly balanced cross-section in 1998, the middle bar shows ANOVA results for a weakly balanced 3-year panel 2011–2013, and the lower bar shows ANOVA results for a weakly balanced 37-year panel. Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

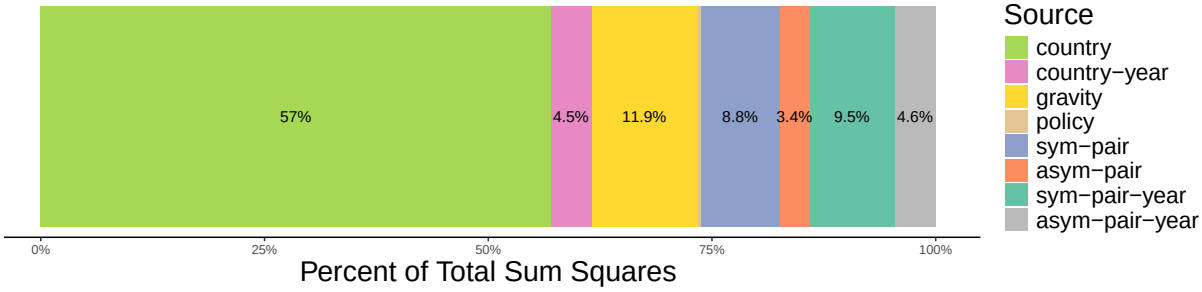
Panel (c) of Figure 1 shows results for the entire unbalanced panel, where the fixed effects capture all dimensions of variation. Figure C.6 shows results for the entire unbalanced panel, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation. Figure C.7 shows results for the entire unbalanced panel, where the first set of fixed effects captures the size and MRT terms in a standard way for gravity estimation, with gravity and policy variables. Panel (a) of Figure 3 shows results for the bilateral variation in the logarithm of trade flows for the entire unbalanced panel. Panel (b) of Figure 3 shows results for the bilateral variation in the logarithm of trade flows for the entire unbalanced panel, with gravity and policy variables.

Figure C.6: ANOVA of the natural logarithm of trade flows, entire panel dataset, years 1980–2016, standard fixed effects to capture size and MRT terms.



Notes: Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

Figure C.7: ANOVA of the natural logarithm of trade flows, entire panel dataset, years 1980–2016, standard fixed effects to capture size and MRT terms, with gravity trade cost proxy variables.



Notes: Percentage shares of less than 2 are not labeled in the plot. Percentage shares of less than 0.0001 are not shown in the plot.

D Details on the Calculations of Shares in the Results Descriptions

Here we give the calculations for the shares reported in Section 4:

- The upper bar of Panel (a) in Figure 1 shows the results of an ANOVA decomposing the variation of the natural logarithm of trade flows for a strongly balanced cross-section in the year 1998 with 53 countries. Among the pair-specific variation, 89% is symmetric $\left(\frac{28.7}{28.7+3.7} \approx 0.89\right)$ and 11% is asymmetric $\left(\frac{3.7}{28.7+3.7} \approx 0.11\right)$.

- Panel (c) in Figure 1 shows the results of an ANOVA of the natural logarithm of trade flows for our entire unbalanced panel in the years 1980–2016 with 223 countries. Expressed as a share of the bilateral variation, about 21% are asymmetric $\left(\frac{3.4+4.6}{21+3.4+9.5+4.6} \approx 0.21\right)$.

Here we give the calculations for the shares reported in Section 5:

- Panel (a) in Figure 3 shows the components for the entire unbalanced panel. In the full panel, 37% of the bilateral variation varies over time $\left(24.7 + 12 \approx 37\right)$.

- Panel (b) of Figure 2 shows that, in weakly balanced subsamples, among the symmetric time-invariant variation—which is essentially the dimension in which all the standard trade cost proxies vary—the standard gravity variables explain 58 to 74% $\left(\frac{47.3}{82} \approx 0.58, \frac{43.2}{74.6} \approx 0.58, \frac{61.6}{82.7} \approx 0.74\right)$.

- In the entire unbalanced panel dataset in Panel (b) of Figure 3, 56% of the symmetric time-invariant variation is explained by the standard gravity variables $\left(\frac{30.8}{54.6} \approx 0.56\right)$.

- Overall, although the trade cost proxy variables together do explain a substantial part of the variation in bilateral trade costs (32 to 67%), they also leave a substantial part unexplained. Of this remaining variation, 33 to 64% is symmetric and time-invariant $\Big(\frac{32.4}{32.4+18} \approx 0.64$, $\frac{29.1}{29.1+14.7+5.6+5.1} \approx 0.53, \frac{15.7}{15.7+4.6+6.8+5.5} \approx 0.48, \frac{22.8}{22.8+8.8+24.6+12} \approx 0.33\Big)$.