GMM over Split Data Sets

Eric Bartelsman[1,2] and Richard Bräuer*[3]

[1]Tinbergen Institute
[2]Vrije Universiteit Amsterdam
[3]Halle Institute for Economic Research

Abstract

GMM estimation currently requires all variables and observations in one data set. This is not feasible in many settings. We propose a technique to recover the GMM estimator if X- and Y-variables are split over multiple data sets or if observations are in separate data sets. Our estimator can also recover individual level regression parameters from aggregate data if shocks are also on aggregate levels (like occupations, industries or regions). Our technique is also faster. Besides simulations, we estimate a cross-European firm level production function from national data sets and argue that the production function differences across countries are likely statistical noise. We also estimate the effect of labor supply shocks on firms' innovation in Prussia from 1895 to 1910, where no firm level data exists. Labor supply increases led to increased patenting by firms without a previous patenting record, but not by incumbents.

*Corresponding author: Richard Bräuer (IWH), Kleine Märkerstraße 8, 06108 Halle (Saale), Richard.braeuer@iwh-halle.de. The paper has benefited greatly from the comments of Jonathan Deist, Filippo di Mauro, Francesco Manaresi, Matthias Mertens, Steffen Müller and the participants of the 2020 CompNet conference. The paper also benefited from the generous help and data provision of Alexander Giebler, Felix Kerstig, Sarah Fritz, Mirko Titze and the CompNet secretariat.

Keywords: Econometrics, GMM, multiple data sets, confidentiality restrictions, firm econometrics
Classification: L11,C01,C13

1 Introduction

The problem of unmergeable or siloed data has long been noted in applied statistics. With the implementation of the EU General Data Protection Legislation (GDPR) in 2016, European administrations are instructed to limit data creation to what is necessary for the original purpose. German administrators in particular have long followed this approach: The Institute for Employment Research (IAB) collects employee biographies and the statistical offices e.g. production data, but it is impossible to link the two because when collection started, firms and employees did not give consent. The problem is also latent in health related fields, where privacy concerns are especially prevalent (Hallock et al., 2021), which often means that patient data has to remain within hospitals. We argue that split data sets do not restrict the researcher as much as previously thought.

This paper presents a new computational strategy for the linear GMM estimator ($\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$). We decompose the estimator into the components of its constituent matrices, which can be computed within siloed data sets. We show analytically that our estimator recovers the GMM parameters exactly even if data is stored across different data sets.[1] This algorithm yields three improvements over conventional methods: First, it recovers $\hat{\boldsymbol{\beta}}$ while information on endogenous variables and outcomes is split over arbitrary unmergeable data sets without extracting individual level data. Second, when exogenous shocks arrive at aggregate levels (e.g. the sector level trade shocks of Autor et al. (2013) or regional shocks like in Danzer et al. (2024)) we can recover $\hat{\boldsymbol{\beta}}$ from aggregate statistics without any access to micro data. Third, our estimator offers speed advantages since it is "embarrassingly parallel", i.e. the computation can be split and relegated to an unrestricted

[1]The code package attached to the paper contains a stata ado-file for distributed linear GMM.

amount of CPUs with minimal overhead.

To contextualize these results, we discuss two simulations and two applications to actual data. In the simulations, we show that our computational strategy recovers the GMM estimator both when variables are in different data sets and when observations are. We also demonstrate that our algorithm is faster than the conventional method and how to recover standard errors.

We apply our estimator to two actual data sets: First, we estimate a European production function. Firm level data is siloed at the national statistical agencies and cannot be merged, which provides a challenge for the study of the common market. We estimate production functions across all European firms as a first step to such an analysis. Second, we recover parameters from aggregate data alone: We study the effect of a labor supply shock on patenting in Prussia: During the "Grain Invasion" (1895-1913), cheap grain from the Americas displaced Prussian agricultural workers. These workers moved to the cities in large numbers. The size of the labor supply shock differed across cities depending on the previous crop specialization of the surrounding countryside (Bräuer and Kersting, 2024). While no firm level data has survived in the historical record, the components that enter the matrices $\boldsymbol{Z}'\boldsymbol{X}$ and $\boldsymbol{Z}'\boldsymbol{Y}$ have been published in the statistical yearbook of Prussia. Our estimator is thus able to recover the firm level estimates from aggregate data.

We are not the first to work with submatrices of the GMM estimator: E.g. MacKinnon et al. (2023) use submatrices to compute jackknife standard errors in a more efficient manner. In medicine, there is an initiative that allows for estimation across samples in a way similar to our second use case (Wolfson et al., 2010). Angrist and Krueger (1992) already suggested to estimate $\boldsymbol{Z}'\boldsymbol{X}$ and $\boldsymbol{Z}'\boldsymbol{Y}$ in completely unrelated data sets. Karim et al. (2024) present an estimator very close to traditional DiD that works if treated and untreated observations cannot be pooled and use it to evaluate Canadian hospitals. We develop a general framework that encompasses these special cases and show the conditions under which any linear GMM estimation can be carried out with unmerged data or even without micro data.

Previously, researchers have bridged across data sets with institutional cooperation, but still rarely could run regressions "across" data sets. The need to maintain a research infrastructure to support these efforts in which all parties are trusted with confidential individual level data also remains a major hurdle for health research (Hallock et al., 2021). In economics, such research networks are also prevalent but face similar issues. Nevertheless, e.g. researchers can currently access European firm level data via the CompNet project which aggregates ex ante decided moments of firm level data (Lopez-Garcia et al., 2014) or the Micro Moments Database (Bartelsman et al., 2017). This is enough to study e.g. European business dynamism (Biondi et al., 2023). Alternatively, the Micro Data Infrastructure (Bartelsman et al., 2023) gives researchers access to actual European firm data under specific conditions. Our algorithm can greatly reduce the costs associated with such research by converting the intermediate steps of a GMM estimation into aggregate data at the appropriate level. Data providers in economics generally have extensive experience with disclosing such data. To further aid them, we provide an in-built customized disclosure check that automatically deletes intermediate results that do not meet specified criteria. Any data provider can thus set their own rules.

Section 2 derives our computational algorithm and discusses the conditions necessary for computation of the point estimates and inference. Section 3 presents examples for the use of the technique. Section 4 concludes.

********** Section on WLS vs. our technique.

Another approach to circumvent the problem of siloed data is to estimate β within each data set and to then export the coefficients, the variance covariance matrix and the sum of residuals. This then allows to compute an inverse-variance weighted mean of the individual β as the overall estimate. Depending on how exactly the variance-covariance matrices and sums of error terms are used, this is equivalent to various special cases of FGLS estimation. FLGS and GMM are identical if the data is known to be homoscedastic, in which case our method would yield the same result.

Our method has two major advantages over this procedure: First, it is more flexible in that it allows $\boldsymbol{Y}$ and $\boldsymbol{X}$ to be in different data sets,

too. In such a case, any FGLS method could not work, since the error terms cannot be computed. Second, FGLS requires that the assumptions about error term structure are actually correct. If the error term structure is misspecified, the FLGS estimator can become biased. As a result, our method of estimating GMM and then using numerical procedures (bootstrap or jackknife) for inference is often preferred even without split data sets. The correctly specified FLGS procedure is more efficient than our estimator only if the individual data sets are large enough so that the variance-covariance matrix can be estimated consistently within each data set. Thus, there may be a place for FLGS if one knows the error structure and the data is large enough so that FLGS can be performed, but not so large that the effciency loss from GMM does not matter. **********

2 Sample bridging GMM

2.1 Point estimation

In this section, we derive the analytical GMM estimator as a function of aggregate statistics out of separate data sets. The standard GMM estimator is defined as $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$.[2] In the joint data set, it would be trivial to compute the result with any statistical software. However, we assume that the data is split over different data sets. For ease of exposition, we stick to 4 data sets A, B, C, D, but the method extends to an arbitrary number of data sets. Both the variables and the observations are spread over these four data sets: A and B contain variables $\boldsymbol{Z}$ and $\boldsymbol{Y}$, C and D contain variables $\boldsymbol{Z}$ and $\boldsymbol{X}$. Observations 1 to K are only in data sets A and C, with the other observations in B and D. Given these definitions of the data sources, we can rearrange the traditional GMM formula:

[2] As usual, $\boldsymbol{Y}$ denotes the vector of the explained variable, $\boldsymbol{X}$ the endogenous explanatory variables and $\boldsymbol{Z}$ the instruments.

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y}) = \begin{bmatrix} \sum_i^N (1 * y_i) \\ \sum_i^N (z_i^1 * y_i) \\ \dots \end{bmatrix} \times \begin{bmatrix} \sum_i^N (1 * 1) & \sum_i^N (1 * x_i^1) & \dots \\ \sum_i^N (z_i^1 * 1) & \sum_i^N (z_i^1 * x_i^1) & \dots \\ \dots & \dots & \dots \end{bmatrix}^{-1} =$$

$$\left[\begin{bmatrix} \sum_{i \in A} y_i \\ \sum_{i \in A} z_i^1 y_i \\ \dots \end{bmatrix} + \begin{bmatrix} \sum_{i \in B} y_i \\ \sum_{i \in B} z_i^1 y_i \\ \dots \end{bmatrix} \right] \times \left[\begin{bmatrix} N_C & \sum_{i \in C} x_i^1 & \dots \\ \sum_{i \in C} z_i^1 & \sum_{i \in C} z_i^1 x_i^1 & \dots \\ \dots & \dots & \dots \end{bmatrix} + \begin{bmatrix} N_D & \sum_{i \in D} x_i^1 & \dots \\ \sum_{i \in D} z_i^1 & \sum_{i \in D} z_i^1 x_i^1 & \dots \\ \dots & \dots & \dots \end{bmatrix} \right]^{-1} \tag{1}$$

β is a function of sums of the product of pairs of variables if estimated conventionally. Our computational strategy fully takes advantage of this: The standard estimator sums up these products of variable pairs directly and inverts $\boldsymbol{Z}'\boldsymbol{X}$. Our strategy is to compute the sums individually within each data set and then use them to replicate the overall estimator. However, this is just a variety of the computational strategy, not in the actual estimation. Our computational strategy exactly recovers $\hat{\boldsymbol{\beta}}$ under all the well understood identifying assumptions of conventional GMM.

Some special cases illuminate the intuition behind our mathematical result: In the first case, data sets are identical in terms of variables, but cannot be merged due to confidentiality. I.e. the researcher would like to "append" the data sets but cannot. In that case, A & C and B & D are the same data set and the technique stops essentially one step short of estimating β within each data set: Instead of reporting $\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y})$ in the two data sets, the estimator reports $\boldsymbol{Z}'\boldsymbol{X}$ and $\boldsymbol{Z}'\boldsymbol{Y}$, from which both the within sample and the across sample $\hat{\boldsymbol{\beta}}$ can be computed. Section 3.2 reports an example of such an application.

In the second case, two data sets cover different variables of the same observations. I.e. the researcher would like to "merge" the data sets but cannot. As long as the instruments $\boldsymbol{Z}$ are present in all data sets, the es-

timation can still proceed. This corresponds to two data sets where A only contains $\boldsymbol{X_A} = \boldsymbol{X}$ and $\boldsymbol{Z_A} = \boldsymbol{Z}$ and data source C only contains $\boldsymbol{Y_C} = \boldsymbol{Y}$ and $\boldsymbol{Z_C} = \boldsymbol{Z}$. Again, the standard computational strategy would not be able to recover the estimator. However, using our method, we are able to circumvent the merge entirely. This is also relevant when merges are legally possible, but technically challenging. E.g. to study the effect of patents it was often necessary to 'record link' patent to firm or individual level data like in Peruzzi et al. (2014); Kline et al. (2019); Akcigit and Goldschlag (2023). Only $\boldsymbol{Z}$ needs to be present in all data sets. Since instruments $\boldsymbol{Z}$ are usually not available in core data sets and have to be constructed by the researcher, this is often not an impediment in practice.

A third special case is if the instruments are identical within data sets, our estimator can recover $\hat{\boldsymbol{\beta}}$ even without access to individual level data. E.g. economic policies are different in different EU NUTS regions, but firm level data is confidential and cannot be appended across regions. Another example of such a data structure is in health economics, where treatment policies differ between hospitals, but patient data is again confidential and cannot leave hospitals. In this case, even aggregate data is enough to compute $\hat{\boldsymbol{\beta}}$: Using $z_i = \bar{z_A}$, we can simplify the sums in our matrix even further to yield $\sum_1^K (z_i^1 * x_i) = z_A * K * \sum_1^K x_i$. This can be computed from aggregate data. Section 3.3 showcases this use case on historical statistical data where firm records did not survive.

A fourth special case is if all data can be assembled to one data set, in which case our computational strategy collapses to the standard. However, even in such a case, there are substantial time savings that come from parallelization as demonstrated by our simulations in section 3.1.

The matrices $\boldsymbol{Z'X}$ and $\boldsymbol{Z'Y}$ do not leak any firm level information: E.g. the first entry simply equals the sum of $\boldsymbol{Y}$ in the data set. This does not run afoul of confidentiality as long as results from regressions within each data set can be published.[3]

[3]This methodology extends to nonlinear GMM, solved numerically. In this case, however, one must compute and store the sum of residuals for all different combinations of possible coefficients in each data set. Summing up these values across all data sets will give the sum of residuals across the whole sample. The coefficient combination with the small-

2.2 Inference

The precision of GMM is usually estimated using the sum of squared residuals $(\boldsymbol{Y} - \beta\boldsymbol{X})'(\boldsymbol{Y} - \beta\boldsymbol{X})$. This presupposes that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are in the same data set, which we are not willing to assume. The exact formula depends on one's assumption about heteroscedasticity, clustering etc. In the special case where all data sets contain all variables (and just cannot be appended in stat lingo), researchers can compute most standard variance estimators by returning to the individual data sets with the estimated coefficients and computing (weighted) sums of error terms.

Outside of this special case, it is not obvious how to compute standard errors, since we cannot compute predicted values. However, non-analytical solutions to inference are still possible. Our computational strategy is especially conducive to clustered standard errors via the jackknife procedure, since we already compute the submatrices necessary for its efficient computation (MacKinnon et al., 2023). We compute standard errors this way when estimating the effect of labor supply shocks on innovation in Prussia, where no firm level data exists. This works because the intuitive strategy of the jackknife is to consecutively drop clusters (not individual observations) and observe the variance of the estimated coefficients. However, this commits the researcher to clustered standard errors at potentially quite high levels of aggregation, depending on the exact setup.

Last, bootstrapping standard errors is also possible. This is true for the bootstrap equivalent of the jackknife presented above. However, bootstrapping within each data set before aggregation can also yield non-clustered standard errors or standard errors clustered at a lower level of aggregation. It only requires that the IDs of the bootstrapped entities are the same in all data sets: To avoid additional assumptions, we want to bootstrap whole observations, which requires coordinating which observations are drawn across data sets without actually sending around ID lists. To do this, we implement a special version of the bootstrap that uses the ID variables as seeds, thereby

est sum of residuals presents the numerical solution to the coefficient vector. However, this is a computationally very expensive procedure.

guaranteeing that a certain observation is drawn the same number of times in every disjoint part of the data set.

Technically, for each bootstrap iteration, we draw a Binomial distributed random variable with $p = \frac{1}{N}$ and $n = N$, where N still denotes the number of observations in the entire data set. This is equivalent to the number of successful independent draws of this observation from the entire data set by the definition of the Binomial distribution. Using the ID of the observation as the seed of the pseudo number generator guarantees that each observation is drawn the same amount of times in each data set.

3 Applications

3.1 Examples: Simulations

To demonstrate our computational strategy, we simulate two different data generating processes, confirm our equivalence result from section 2 in practice and display the main applications. We also show the speed advantage of our strategy.

First, we simulate a big data setting, an application which has become an important part of the economy (Veldkamp and Chung, 2024). Consider a firm monitoring machine failure from the readings of a set of 5 sensors reporting temperature fluctuations each second at various places in the machines. Such data quickly accumulates: After one year, this creates a data set of 3.2 billion observations per machine. However, the algorithm we propose can achieve this regression in a reasonable time by computing the sums constituting $\boldsymbol{Z_I}'\boldsymbol{Z_I}$ and $\boldsymbol{Z_I}'\boldsymbol{Y_I}$ for every machine on a separate CPU. In a practical application, one could even compute these running sums at measurement and skip the need for data transfer and a data set entirely (distributed computing). Specifically, we rearrange equation (1) to yield

$$\hat{\boldsymbol{\beta}} = \left(\sum_{I} \begin{bmatrix} \sum_{i \in I} y_i \\ \sum_{i \in I} z_i^1 y_i \\ \dots \end{bmatrix} \right) \times \left(\sum_{I} \begin{bmatrix} \sum_{i \in I} 1 & \sum_{i \in I} z_i^1 & \dots \\ \sum_{i \in I} z_i^1 & \sum_{i \in I} z_i^1 z_i^1 & \dots \\ \dots & \dots & \dots \end{bmatrix} \right)^{-1} \tag{2}$$

where I indexes the different data sets (= machines), each of which contains N_I observations indexed by i. Table 1 reports the traditional OLS result and the identical estimate obtained from our own algorithm, together with computing times and data transfer requirements using both methods. To attain the theoretically possible 75% runtime reduction with four machines, one would have to leave the stata programming language to avoid as much overhead as possible. While economists do not (yet) run regressions of such magnitudes often, big data applications have gained relevance in economics. The ability of our algorithm to bring an unbounded number of cores to bear on such problems can already create substantial time savings for today's applications.

Second, we simulate a scenario with $\boldsymbol{Y}$ and $\boldsymbol{X}$ in different data sets with the researcher unable to perform a merge. Consider a researcher trying to merge patent data to economic variables. Patent data only contains string names instead of any actual IDs. There is extensive work to disambiguate names, add location variables from patent texts (Toole et al., 2021; Bergeaud and Cyril, 2022) and use these variables to merge patents to other data (Peruzzi et al., 2014) via string matching. We argue that for some questions, this is actually unnecessary. Specifically, we simulate a change of firms' R&D subsidy regime ($\boldsymbol{Z}$), affecting their actual subsidy uptake ($\boldsymbol{X}$) and patenting outcomes ($\boldsymbol{Y}$). This is reminiscent of the data situation in Germany, where EU subsidy generosity quasi-exogenously varies at the county level (Brachert et al., 2019) and firm level subsidy data exists (Brachert et al., 2018). Merging this information to statistical data sets is hard, especially since administrative sources cannot be combined with each other (Fritsch et al., 2020). However, the county level change in subsidy policy is trivial to merge to both patent and statistical office data, so we rearrange equation

Table 1: Estimations from Simulations

	Machine Failure		Patent Merge	
	OLS	B-GMM	IV	B-GMM
$X1$	0.600***	0.600***		
	(0.004)	(0.006)		
$X2$	0.607***	0.607***		
	(0.004)	(0.004)		
$X3$	0.593***	0.593***		
	(0.004)	(0.006)		
$X4$	0.602***	0.602***		
	(0.004)	(0.005)		
$X5$	0.599***	0.599***		
	(0.004)	(0.006)		
lag R&D expenses			1.001***	1.001***
			(0.003)	(0.003)
R^2	0.00	0.00	0.17	0.17

Notes: Result of both simulation exercises. The columns OLS/IV refer to the estimate from the conventional estimator, B-GMM refers to our new estimate. Heteroscedasticity robust standard errors in column 1, 3 and 4. Cluster robust (jackknife) standard error at the machine level in column 3 to demonstrate both versions of inference. Significance: *10 %, **5 %, ***1 %.

(1) to yield

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{X})^{(-1)}(\boldsymbol{Z}'\boldsymbol{Y}) = \begin{bmatrix} \sum y_i \\ \sum z_i^1 y_i \end{bmatrix} \times \begin{bmatrix} N & \sum x_i^1 \\ \sum z_i^1 & \sum z_i^1 x_i^1 \end{bmatrix}^{-1} \tag{3}$$

which allows us to estimate the IV coefficient without merging $\boldsymbol{Y}$ and $\boldsymbol{X}$ into the same data set. Table 1 again reports the traditional IV result together with the identical estimate computed from our algorithm.

3.2 Use Case 1: European Production Function Estimation

A representative data set of European firms does not exist. Researchers have to either restrict their analysis to singular countries or use BvD data for Europe (AMADEUS, ORBIS) which does not cover new or privately owned firms well. To alleviate this problem, the CompNet network unifies variable definitions across European firm level data sets and publishes meso aggregated data of a host of indicators. Nevertheless, regressions with observations from different countries are not possible. Neither are they feasible with the Micro Data Infrastructure (Bartelsman et al., 2023), which allows direct access to the firm level data sets of some countries.

This is an issue for productivity analysis insofar as firms' productivity can only be confirmed if it is measured as the residual from the same production function. We can solve this issue by estimating a cross-country production function for European two digit industries. We ran our codes in the experimental module of CompNet 2019, which 14 European countries consented to run, namely Belgium, Croatia, Czechia[4], Denmark, Finland, Italy, Lithuania, Netherlands, Portugal, Romania, Slovenia, Spain, Sweden and Switzerland. For a detailed description of the underlying data sets, their coverage and sampling processes, we refer the reader to Lopez-Garcia et al. (2014).

To recover a measure of firm productivity (i.e. TFPR), we estimate revenue as a Cobb-Douglas function of labor, capital and intermediate inputs. After taking logs, one estimates

$$q_{it} = \beta^l l_{it} + \beta^k k_{it} + \beta^m m_{it} + \omega_{it} + \varepsilon_{it}, \tag{4}$$

where l_{it}, k_{it} and m_{it} denote log labor, intermediates and capital, ε_{it} denotes random measurement error or short term shocks and firm log TFP is ω_{it}. Equation (4) is presumably endogenous, most importantly because input choices l_{it}, k_{it} and m_{it} are themselves decisions that the firm makes after observing its own productivity ω_{it}, which is part of the error term. To cir-

[4]The results below do not contain Czechia, where code execution stopped before the results could be constructed.

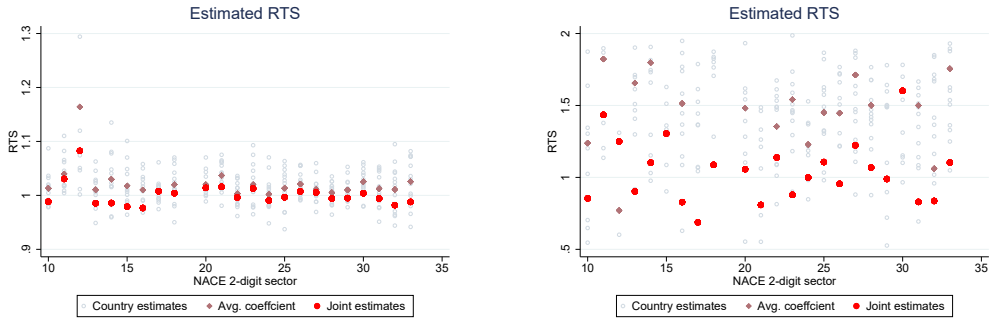
cumvent this endogeneity problem, researchers routinely use the Wooldridge (2009) estimator, which assumes that firms can no longer change their static inputs (labor and capital in our case) by the time they observe this year's productivity. This assumption makes it possible to construct and estimate a control function that captures the response of firms to productivity shocks. We employ this technique as an alternative specification to deal with endogeneity concerns, which are however not at the center of our discussion. More evolved control function estimators (Ackerberg et al., 2015; De Loecker et al., 2016) are solved numerically, which would still in principle be possible to estimate by extracting the sum of residuals for different parameterizations in each country and then minimizing the sum across countries, but this technique placed undue burdens on the data providers, given that actual estimates are usually very close to the simple versions.

Figure 1 compares both the returns to scale and the labor coefficients of the joint estimation and the within country estimates. They demonstrate how volatile the within country estimations can be. Generally, the control function approach yields more volatile estimates than OLS, a well known problem that is again exacerbated by small sample sizes. Even the joint control function estimation still has outliers, e.g. "Beverages" and the "Other transportation equipment" with returns to scale of roughly 1.5. The average returns to scale of slightly above 1 are to be expected since concentration measures for the European market are trending upwards (Bighelli et al., 2022). This volatility could either be because the different countries have legitimately different production functions or it could be random noise.

To further explore the differences between the joint and the within country estimates, we analyze which countries have systematically different results. Specifically, we use 296 sector-country level estimation results from equation (4) and estimate

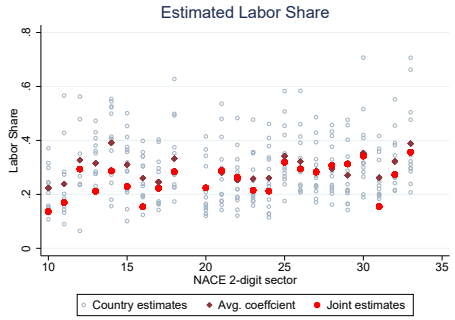
$$\beta_c - \beta_j = \gamma_c + \rho_s + u_{c,s} \tag{5}$$

where β_c is the country specific regression coefficient, β_j is the jointly estimated coefficient, γ_c is a country fixed effect and ρ_s is a sector specific fixed

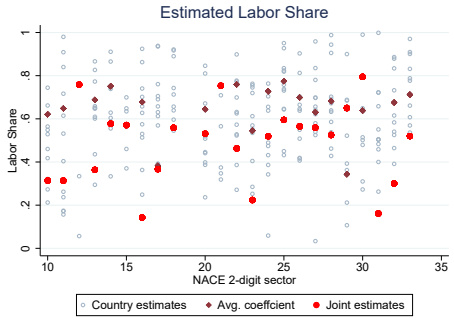


(a) OLS

(b) Control Function



(c) OLS



(d) Control Function

Figure 1: Labor coefficient and returns to scale when estimating a standard Cobb-Douglas revenue production function using OLS or a control function approach. Red dots give the returns to scale estimated in the cross country data while blue circles give the individual country estimates. Diamonds denote the (weighted) average of country coefficients. Note the different scaling on the OLS and the control function graph. Data: CompNet

effect. Note the absence of a constant. Table 2 reports the γ_c from this regression. If countries' production functions differ systematically because of e.g. a better educated workforce in one country, we would expect the labor coefficient to be larger in all sectors of the economy. However, if the deviations are largely random measurement error, they cannot be explained by country fixed effects.

Overall, we find the latter: Out of the 13 countries, only Sweden, Denmark, Lithuania and Croatia have systematic deviations from the joint OLS estimate. These are potentially problematic insofar as they might be interpreted as these countries having systematically different production functions. Interestingly, all of them have substantially, arguably even implausibly higher labor coefficients. E.g. Sweden allegedly uses twice as much labor to produce the same goods as the average of the other countries. Given the very high labor costs and capital stock of the country, that seems quite unlikely. The only other country with such high labor coefficients is Croatia, but it has totally different conditions. Of these four, only Croatia has a similar deviation in the control function approach.[5] All of these outlier countries have small data sets. All in all, we do not think that these differences can be interpreted as different, country specific production functions. Instead we see them as random errors caused by the small sample size in some countries.

3.3 Use Case 2: Estimating the Effect of a Labor Supply Shock on Innovation from destroyed Prussian firm data

Firm level data from Germany does not exist prior to 1974. Nevertheless, we use our computational strategy to study the effect of a labor supply shock on innovation decisions of Prussian firms from 1895-1910. This is possible since labor supply shocks are by design at the level of the local labor market,

[5]We have to exclude roughly 20% of country-sector production function estimates because they are obviously unusable, i.e. they have negative coefficients or returns to scale above 2.

Table 2: Country Differences in Production Functions

	RTS (OLS) (1)	α^l (OLS) (2)	RTS (CC) (3)	α^l (CC) (4)
BELGIUM	0.000	0.000	0.000	0.000
	(.)	(.)	(.)	(.)
CROATIA	0.013	0.265***	0.001	0.372***
	(0.016)	(0.038)	(0.101)	(0.130)
DENMARK	0.039**	0.213***	0.052	-0.013
	(0.016)	(0.038)	(0.101)	(0.130)
FINLAND	-0.020	0.036	0.071	0.071
	(0.016)	(0.038)	(0.103)	(0.133)
ITALY	0.014	0.015	0.172	0.034
	(0.016)	(0.038)	(0.106)	(0.136)
LITHUANIA	0.082***	0.144***	-0.060	-0.148
	(0.016)	(0.038)	(0.104)	(0.134)
NETHERLANDS	-0.021	-0.008	0.077	-0.072
	(0.016)	(0.039)	(0.101)	(0.130)
PORTUGAL	0.033**	0.091**	0.151	0.243
	(0.016)	(0.038)	(0.118)	(0.152)
ROMANIA	0.034**	0.001	0.002	-0.155
	(0.016)	(0.039)	(0.106)	(0.137)
SLOVENIA	0.005	0.045	0.195*	0.119
	(0.016)	(0.039)	(0.108)	(0.139)
SPAIN	0.026	0.086**	0.344***	-0.030
	(0.016)	(0.038)	(0.108)	(0.140)
SWEDEN	-0.008	0.283***	-0.209**	-0.060
	(0.016)	(0.038)	(0.098)	(0.127)
SWITZERLAND	0.003	0.084**	-0.051	0.201
	(0.016)	(0.038)	(0.110)	(0.142)
N. Obs	296	296	231	231
R sq	0.310	0.411	0.631	0.508

Notes: Results obtained from equation (5). Columns report differences in returns to scale (RTS) and the labor coefficient α^l for the OLS and the control function approach. Belgium is used as baseline. Output from sector level fixed effects is not shown. The number of observations changes because all sector-country production functions with returns to scale more than 100% higher than the joint estimate were dropped. Data: CompNet
Significance: *10 %, **5 %, ***1 %.

which we proxy with the county. The Prussian Statistical Office published the number of firms per industry sector and their employment at a county level in 1882, 1895 and 1910, the years of the imperial census.

We use this data to study the effect of the "Grain Invasion" – the inflow of cheap American grain into Europe – on industrial firms and their innovation decisions in Prussia. Bräuer and Kersting (2024) report substantial effects of this trade shock on agricultural counties following the methodology of Autor et al. (2013), but a different response compared to contemporary shocks: Workers in Prussia moved from affected counties into the cities, which accelerated structural change substantially. Since 76% of all patents came from these big cities even prior to the shock, we focus on the innovation there, not in the rural counties originally hit by the "Grain Invasion". Even patents concerning agricultural techniques are often developed in the agricultural colleges in the cities and thus not easily attributable to any specific shock.

The innovation effect of trade shocks is usually studied in firm level data (Bloom et al., 2016; Autor et al., 2020; Bräuer et al., 2023), while the effects of labor supply shocks are often estimated with aggregate or cross sectional data (Danzer et al., 2024). Even though no firm level data exists, we can estimate both versions and shed light on the difference between the two. We estimate

$$\Delta_{1895,1910} p_f = g_{1895,1910}(L_c) + \beta_x X_c + u_f \tag{6}$$

where $\Delta_{1895,1910} p_f$ denotes the change in the number of patents of firm f and $g_{1895,1910}(L_c)$ denotes the growth rate of the county population, both between 1895 and 1910. X_c denotes a set of county level control variables, specifically the share of agricultural employment, the share of large estates among farms, the distance to the next metropolis and the installed stock of steam engines (measured in horsepower in 1875). We follow Bräuer and Kersting (2024) in this, who argue these variables proxy the growth potential of the counties apart from trade or labor supply shocks. We also use their instruments: the import shock due to grain imports per county and the indirect shocks due to immigration from shocked counties constructed from

province and county migration data prior to the shock. We refer the reader to their paper for a detailed description of the construction of these variables. Our estimation strategy leaves us with a $\boldsymbol{Z}$ vector of four control variables and two instruments (direct and indirect trade shock), and an $\boldsymbol{X}$ vector of the same four exogenous regressors and county population growth rate L_c. To compute the results, we rearrange equation (1) to

$$\hat{\boldsymbol{\beta}} = \left(\sum_I \begin{bmatrix} N_I \bar{y}_I \\ \bar{z}_I^1 N_I \bar{y}_I \\ \dots \end{bmatrix} \right) \times \left(\sum_I \begin{bmatrix} N_I & \bar{x}_I^1 N_I & \dots \\ \bar{z}_I^1 N_I & \bar{z}_I^1 \bar{x}_I^1 N_I & \dots \\ \dots & \dots & \dots \end{bmatrix} \right)^{-1} \tag{7}$$

Note that all instruments are constant at the level of each data set indexed I (i.e. county). This allows us to reformulate the regression coefficients not just as a function of the underlying micro data, but also as a function of very commonly published aggregate statistics (means and observations counts). Table 3 reports the results from this estimation technique and the regression of aggregates. We have dropped outliers, singleton counties and counties without information on the number of firms to arrive at our estimation sample.

Column (1) of table 3 reports the aggregate regression in percentages: The effect of increasing the population by 1% is a 1.8% increase in patenting. Patents growing faster than population is surprising insofar as the immigrants were low education farm hands, so their own contribution to patents is likely minimal. However, it fits into the overall picture of much better adjustment in Prussia: Instead of suffering the losses in affected counties, workers moved to the cities while cities' industries absorbed them without income per capita losses. The rise in patenting might be one of the channels through which cities could do this. Column (2) estimates the same relationship, but uses the first difference in the number of patents and the population increases. This is unusual, but required for firm level analysis, since first differences add up to the aggregate first difference (which is reported in the statistical yearbooks), while individual percentage increases do not add up and thus cannot be recovered. Column (3) re-estimates column (2) with our own estimator, but while setting the number of firms in each county N_I to one. This recreates

Table 3: The Effect of a Labor Supply Shock on Innovation in Prussia 1895-1910

	Aggregate		Firm Level		
	IV	IV	B-GMM	G-GMM	B-GMM
Pop. growth (%)	3.243**				
	(1.52)				
Pop. increase (1000)		0.27***	0.27***	0.29***	-0.05***
		(0.03)	(0.03)	(0.03)	(0.02)
share agri.	2.78	0.622	0.622	-13.3	36.25
	(16.06)	(45.54)	(47.10)	46.06	(25.93)
share big farms	0.000	0.17	0.17	0.02	-0.40***
	(0.05)	(0.21)	(0.11)	(0.20)	(.09)
horsepower per worker	-1.75	-22.66*	-22.66	-25.36	21.59***
	(4.89)	(12.00)	(15.52)	(30.73)	(8.43)
distance	0.00	-0.13	-0.13*	-.14**	.20***
	(0.02)	(0.10)	(0.55)	(0.10)	(.04)
R^2	0.00	0.67	0.46	-	-
counties	52	52	52	52	52
Observations	52	52	52	437140	437140

Notes: Results obtained from equation (7), standard errors clustered at the district level. Column one and two give the effect of a population increase on patenting in percent and number of people estimated on the county level. Column 3 reports the results from estimating column 2 with our new algorithm, but setting the number of firms in each county to 1. Evidently, both point estimates are equivalent. Standard errors are clustered at the district level and derived analytically for column (2) and with the jackknife for column (3). This leads to different error bands due to the small number of clusters. Column 4 reports the firm level results, effectively a reweighting of column (3). Column 5 reports the incumbents-only results, a specification also often used in firm level analysis. To ensure a causal estimation, we use the indirect trade shock of Bräuer and Kersting (2024): An inflow of rural goods that caused urbanization and hit some cities' countrysides harder than others. Data: Bräuer and Kersting (2024)
Significance: *10 %, **5 %, ***1 %.

the estimation from column (2). Column (4) is the firm level regression. The firm level estimate is a reweighted aggregate regression because the exogenous variables vary at the aggregate level. Column (5) restricts the first difference in patenting to incumbent firms. It is the only estimate that produces a substantially different result from the other specifications. It seems that the increase in patenting is mainly driven by newly entering firms, while incumbents behave much more like modern firms or counties (Danzer et al., 2024).

While patenting for individual firms can be recovered from the patent data, the statistical sources for Prussia only report aggregate outcomes. This is not in principle an obstacle to estimation, but it restricts the specifications that can be estimated to those where variables "add up" to their aggregate level counterparts. As in this application, researchers might be restricted to first difference instead of e.g. log-linear specifications. Column (4) and (5) also suffer from the fact that Prussia did not separately report firm entry, so the number of firms in each county is the same in (4) and (5) even though (5) aims to estimate with incumbents only. However, in contemporary applications, this is less of a problem (e.g. the US census bureau reports aggregate statistics by firm birth year). Individual level X-variables are of course possible, but again might be difficult to implement in practice if lagged values are required. These are not often reported as aggregate statistics, though this is changing: E.g. the Business Dynamics Statistics by the US census bureau report statistics per cohort of firms that would allow the construction of lagged aggregate values.

4 Conclusion

This paper shows the viability of a new algorithm that can estimate GMM in dispersed, unmergeable data sets. This method can account for X- and Y-variables being stored in different data sets and observations being dispersed over different data sets. The algorithm can cope with both problems simultaneously. We demonstrate the substantial time savings possible in the special case of one data set, even within the stata language. We also show

how to recover individual level regressions from aggregate data in special cases where exogenous shocks are on the same level as aggregate information (e.g. counties or sectors). While our technique is much more permissive than the standard computational strategy, it requires that all instruments and exogenous regressors are present in all data sets.

In two simulations, we demonstrate that our computational strategy recovers the identical parameters that a GMM in the pooled data set would have yielded, if it were possible. We also show our technique generates substantial savings in both computation time and potentially data storage needs.

We also present two applications: First, we estimate a firm level production function across Europe, even though no pooled representative data set of European firms exists. For this, we use access to 14 European data sets of statistical offices and central banks via the CompNet network. We conclude that the production function differences between especially small European countries are likely just mismeasurement and do not capture actually different production technologies. Second, we estimate the effect of a labor supply shock on Prussian firms' innovation and outcomes, despite no firm level data existing. We find that labor shocks invigorated patenting in Prussia on the aggregate, driven by first-time patenting firms. This is in stark contrast to today, where innovation moves towards automation but the number of patents does not increase substantially (Danzer et al., 2024).

5 Bibliography

H. Hallock, S. E. Marshall, P. A. C. 't Hoen, J. F. Nygård, B. Hoorne, C. Fox, S. Alagaratnam, Federated Networks for Distributed Analysis of Health Data, Frontiers in Public Health 9 (2021). Publisher: Frontiers.

D. H. Autor, D. Dorn, G. H. Hanson, The China Syndrome: Local Labor Market Effects of Import Competition in the United States, American Economic Review 103 (2013) 2121–2168.

A. M. Danzer, C. Feuerbaum, F. Gaessler, Labor supply and automation in-

novation: Evidence from an allocation policy, Journal of Public Economics 235 (2024) 105136.

R. Bräuer, F. Kersting, Trade Shocks, Labour Markets and Migration in the First Globalisation, The Economic Journal 134 (2024) 135–164.

J. G. MacKinnon, M. O. Nielsen, M. D. Webb, Leverage, influence, and the jackknife in clustered regression models: Reliable inference using summclust, Stata Journal 23 (2023) 942–982. Publisher: StataCorp LP.

M. Wolfson, S. E. Wallace, N. Masca, G. Rowe, N. A. Sheehan, V. Ferretti, P. LaFlamme, M. D. Tobin, J. Macleod, J. Little, I. Fortier, B. M. Knoppers, P. R. Burton, DataSHIELD: resolving a conflict in contemporary bioscience–performing a pooled analysis of individual-level data without sharing the data, International Journal of Epidemiology 39 (2010) 1372–1382.

J. D. Angrist, A. B. Krueger, The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples, Journal of the American Statistical Association 87 (1992) 328–336. Publisher: [American Statistical Association, Taylor & Francis, Ltd.].

S. Karim, M. D. Webb, N. Austin, E. Strumpf, Difference-in-Differences with Unpoolable Data, 2024. ArXiv:2403.15910 [econ].

P. Lopez-Garcia, F. Di Mauro, N. Benatti, C. Angeloni, C. Altomonte, M. Bugamelli, L. D'Aurizio, G. Barba Navaretti, E. Forlani, S. Rossetti, D. Zurlo, A. Berthou, C. Sandoz-Dit-Bragard, E. Dhyne, J. Amador, L. D. Opromolla, A. C. Soares, B. M. Chiriacescu, A.-M. Cazacu, T. Lalinsky, E. Biewen, S. Blank, P. Meinen, J. Hagemejer, P. Tello, A. Rodríguez-Caloca, U. Čede, K. Galuscak, J. Meriküll, P. Harasztosi, Micro-Based Evidence of EU Competitiveness: The CompNet Database, SSRN Electronic Journal (2014).

E. Bartelsman, E. Hagsten, M. Polder, Micro Moments Database for Cross-Country Analysis of ICT, Innovation, and Economic Outcomes, Working Paper 17-003/IV, Tinbergen Institute Discussion Paper, 2017.

F. Biondi, S. Inferrera, M. Mertens, J. Miranda, Declining Business Dynamism in Europe: The Role of Shocks, Market Power, and Technology, Jena Economics Research Papers (2023). Number: 2023-011 Publisher: Friedrich-Schiller-University Jena.

E. Bartelsman, I. Dorn, M. Haelbig, A. Z. Mattioli, Micro Data Infrastructure: Documentation, MDI technical reports (2023).

M. Peruzzi, G. Zachmann, R. Veugelers, Remerge: Regression Based Record Linkage With An Application To PATSTAT, Bruegel Working Paper 10 (2014).

P. Kline, N. Petkova, H. Williams, O. Zidar, Who Profits from Patents? Rent-Sharing at Innovative Firms, The Quarterly Journal of Economics 134 (2019) 1343–1404.

U. Akcigit, N. Goldschlag, Measuring the Characteristics and Employment Dynamics of U.S. Inventors, 2023.

L. Veldkamp, C. Chung, Data and the Aggregate Economy, Journal of Economic Literature 62 (2024) 458–484.

A. Toole, C. Jones, S. Madhavan, PatentsView: An Open Data Platform to Advance Science and Technology Policy, USPTO Ec. WP 2021-1, 2021.

A. Bergeaud, V. Cyril, PatentCity: a dataset to study the location of patents since the 19th century, 2022.

M. Brachert, E. Dettmann, M. Titze, The regional effects of a place-based policy – Causal evidence from Germany, Regional Science and Urban Economics 79 (2019) 103483.

M. Brachert, A. Giebler, G. Heimpold, M. Titze, D. Urban-Thielicke, IWH-Subventionsdatenbank: Mikrodaten zu Programmen direkter Unternehmenssubventionen in Deutschland. Datendokumentation, Research Report 2/2018, IWH Technical Reports, 2018.

M. Fritsch, M. Titze, M. Piontek, Identifying cooperation for innovation—a comparison of data sources, Industry and Innovation 27 (2020) 630–659. Publisher: Routledge _eprint: https://doi.org/10.1080/13662716.2019.1650253.

J. M. Wooldridge, On estimating firm-level production functions using proxy variables to control for unobservables, Economics Letters 104 (2009) 112–114.

D. A. Ackerberg, K. Caves, G. Frazer, Identification Properties of Recent Production Function Estimators, Econometrica 83 (2015) 2411–2451. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA13408.

J. De Loecker, P. K. Goldberg, A. K. Khandelwal, N. Pavcnik, Prices, Markups, and Trade Reform, Econometrica 84 (2016) 445–510. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.3982/ECTA11042.

T. Bighelli, F. di Mauro, M. J. Melitz, M. Mertens, European Firm Concentration and Aggregate Productivity, Journal of the European Economic Association (2022).

N. Bloom, M. Draca, J. Van Reenen, Trade Induced Technical Change? The Impact of Chinese Imports on Innovation, IT and Productivity, The Review of Economic Studies 83 (2016) 87–117.

D. Autor, D. Dorn, G. H. Hanson, G. Pisano, P. Shu, Foreign Competition and Domestic Innovation: Evidence from US Patents, American Economic Review: Insights 2 (2020) 357–374.

R. Bräuer, M. Mertens, V. Slavtchev, Import competition and firm productivity: Evidence from German manufacturing, The World Economy 46 (2023) 2285–2305. Publisher: John Wiley & Sons, Ltd.

Appendices

A Results of production function estimation

We estimate sector level production elasticities across all NACE 2-digit industries from 4. Table 4 reports the results for all manufacturing sectors, using simple OLS.

The most mechanical benefit of using the CompNet database in this way is that we can estimate these coefficients with 1.9 million observations, which yields a high precision estimate even for small sectors of the economy. One interesting finding is that using this mass of firms, estimated returns to scale are very close to 1 using OLS. The only exception is the tobacco industry which even in all our countries combined only consists of 500+ firms. Our method uncovers coefficients that are broadly in line with the literature at large, with the intermediate goods coefficient between $\frac{2}{3}$ and $\frac{3}{4}$. This is a byproduct of using OLS estimates, since firms adjust intermediate inputs fastest when productivity changes thus this coefficient has some reverse causality issues. This lowers the labor and capital coefficients as well. To adjust for this problem, we employ an instrumental variable approach as described in section 3.2. Large data sets alone cannot rectify this problem.

Table 4: Output Elasticities by Sector (OLS)

Sector		# obs (1)	m (2)	l (3)	c (4)	RTS (5)
10	Food products	201,173	0.82	0.14	0.04	0.99
11	Beverages	23,235	0.84	0.17	0.02	1.03
12	Tobacco	536	0.76	0.29	0.03	1.08
13	Textiles	63,263	0.75	0.21	0.02	0.99
14	Apparel, dressing etc.	81,592	0.68	0.29	0.02	0.99
15	Leather & leather products	49,787	0.71	0.23	0.04	0.98
16	Wood & wood products	10,3537	0.78	0.15	0.04	0.98
17	Pulp & paper products	30,247	0.75	0.22	0.03	1.01
18	Printing & Replication	96,031	0.69	0.28	0.03	1.00
20	Chemical products	49,454	0.76	0.22	0.03	1.01
21	Pharmaceuticals	8,357	0.72	0.29	0.01	1.02
22	Rubber & Plastics	90,841	0.71	0.26	0.03	1.00
23	Glas & Ceramics	94,234	0.77	0.22	0.03	1.01
24	Basic metals	26,007	0.75	0.21	0.02	0.99
25	Fabricated metal products	362,388	0.63	0.32	0.05	0.99
26	Electronics & optics	49,791	0.70	0.29	0.01	1.00
27	Electrical equip.	52,080	0.71	0.28	0.01	1.01
28	Machinery & equipment	166,997	0.66	0.31	0.02	0.99
29	Motor vehicles	31,055	0.67	0.31	0.01	0.99
30	Other transport equipment	19,842	0.65	0.34	0.02	1.00
31	Furniture manufacturing	92,345	0.81	0.15	0.03	0.99
32	Other manufacturing	68,649	0.68	0.27	0.03	0.98
33	Repair & installation	109,881	0.59	0.35	0.04	0.99

Notes: Results obtained from equation (4) estimated as OLS per sector. Columns 1-5 report the number of observations, the output elasticities for intermediate, labor, and capital inputs and the returns to scale. Clustering at the firm level. Significance: *10 %, **5 %, ***1 %.