Aggregravity: Estimating gravity models from aggregate data

Badinger, Harald; Crespo Cuaresma, Jesus

Published in:
Applied Economics

DOI:
10.1080/00036846.2014.1002903

Published: 01/01/2015

Document Version
Peer reviewed version

Aggregravity: Estimating Gravity Models from Aggregate Data[*]

Harald Badinger[†]
Jesus Crespo Cuaresma[‡]

Abstract

This paper considers alternative methods to estimate econometric models based on bilateral data when only aggregate information on the dependent variable is available. Such methods can be used to obtain an indication of the sign and magnitude of bilateral model parameters and, more importantly, to decompose aggregate into bilateral data, which can then be used as proxy variables in further empirical analysis. We perform a Monte Carlo study and carry out a simple real world application using intra-EU trade and capital flows, showing that the methods considered work reasonably well and are worthwhile being considered in the absence of bilateral data.

Keywords: Aggregation, gravity equations.
JEL Classifications: C13, F14, F17.

[*] The authors would like to thank an anonymous referee for helpful comments.

[†] Corresponding author: Department of Economics, Vienna University of Economics and Business (WU) and Austrian Institute for Economic Research (WIFO). Address: Welthandelsplatz 1, 1020 Vienna (Austria). Email: harald.badinger@wu.ac.at

[‡] Department of Economics, Vienna University of Economics and Business (WU); Wittgenstein Centre for Demography and Global Human Capital (WIC); World Population Program, International Institute of Applied Systems Analysis (IIASA) and Austrian Institute for Economic Research (WIFO). Address: Welthandelsplatz 1, 1020 Vienna (Austria). Email: jcrespo@wu.ac.at.

1 Introduction

In many empirical economic applications, information on bilateral relationships between economic units is desired but only aggregated data exists for the variable of interest. This may be directly related to the low aggregation level considered, e.g., when bilateral subnational data are required for a variable that is only available at the country level. Alternatively, this may occur for specific variables of interest at the same aggregation level, e.g., for balance of payments data. While bilateral trade data are readily available, this is not the case for bilateral capital account data, let alone its sub-accounts.

On the other hand, at least for several variables of interest, data on variables explaining these bilateral relationships can be obtained. The gravity model is a leading case in point. As shown by Frankel and Romer (1999), bilateral and aggregate geographical information on countries (which is readily available) can explain a large share of the variation in bilateral trade across countries. The same argument applies to capital flows such as foreign direct or portfolio investment (Sarisoy Guerin, 2006) as well as migration flows/stocks (Abel, 2013). This suggests that observed data on exogenous bilateral variables could be used to generate reasonably close approximations to country-pair specific bilateral data on the unobserved variables of interest.

In an increasingly integrated world economy where linkages of various kinds, both at the country and the regional level, are becoming increasingly important, the lack of bilateral data on economically relevant linkages across countries (or, more generally, economic units) is a major shortcoming. Hence, when bilateral data (which would obviously be the first best solution) are not available, decomposing aggregate data into bilateral relationships is a topic of obvious interest for research, at least as long as bilateral data remain unavailable. Accordingly, there is evidence that gravity models work well not only at the country level but also at the regional level (see for instance Mitze et al., 2010), for which data on relevant explanatory geographical and socio-demographic variables are readily available. Approaches that can be used to generate reasonably close approximations to region-pair specific bilateral data on variables that are only available at a country level (or a higher regional aggregation level) appear thus important in this context.

This contribution aims assessing approximation methods to estimate (disaggregated) bilateral models when only (aggregate) country-specific data is available on the dependent variable, as it is the case the aforementioned contexts. These methods can thus be used to overcome the current lack of bilateral data in some economic applications such as modelling financial flows between countries. This allows to generate approximations of disaggregated information from aggregate data based on theoretically founded and empirically established models.

Our analysis is related to the literature on non-linearly aggregated models. Lee et al. (1993) and Granger and Lee (1999), for instance, consider the effect of aggregating series which are generated nonlinearly to form aggregated cross-sectional or temporal data. Most of the contributions in this area deal with the effects of aggregating high-frequency data to lowerfrequency observations. The majority of the existing studies concentrate on assessing the differences in the time series characteristics of the aggregate as compared to the individual series that form the low-frequency dataset (see Silvestrini and Veredas, 2008, for a survey on the issue of temporal aggregation of time series specifications). Our contribution assesses the question of aggregation from the opposite perspective and focuses on cross-sectional data. The general question we are thus trying to answer is: when facing cross-sectional data that has been created by nonlinearly aggregating variables measured at a higher degree of detail, how can we get reliable estimates of the building blocks of the aggregated observations? Conceptually, our aim is not very different from that in Proietti (2006), that is, to obtain data at higher frequency than those available (e.g., quarterly from annual data) under the assumption that these have been aggregated nonlinearly.

In line with this strand of the literature, the present paper does not aim at providing a substitute for estimation and inference in bilateral models. Rather, the methods considered in the present paper aim at approximating (unobserved) bilateral data as close a possible, which can be then be used in further empirical analyses. We consider two alternative approaches based on maximum likelihood (ML) estimation and linearized generalized least squares (GLS) methods. Our Monte Carlo results indicate that both methods perform relatively well at obtaining point estimates of the parameters of the bilateral model using aggregated information. We carry out a simple empirical application in which we estimate elasticities for bilateral models of exports, asset claims, foreign direct investment and portfolio investment using aggregated data.

The remainder of the paper is organized as follows. Section 2 sets up the econometric framework and outlines two approximation methods to estimate bilateral models from aggregate data. Sections 3 and 4 provide a simple Monte Carlo exercise and a real world data application to assess the performance of the approximation methods. The final section 5 summarizes the results and concludes.

2 Econometric Framework

Consider a linear model for (unobserved) bilateral data corresponding to N individuals,

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \tag{1}$$

where $\mathbf{y}$ is an N^2-dimensional vector, $\mathbf{X}$ is an $N^2 \times K$ known matrix of explanatory variables, $\boldsymbol{\beta}$ is a $K \times 1$-dimensional vector and $\boldsymbol{\varepsilon}$ is an $N^2 \times 1$-dimensional error term, i.e., $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I}_{N^2})$. Let the observed (aggregate) variable be given by an N-dimensional vector $\mathbf{Y}$ such that each element of the vector is given by $\mathbf{Y}_i = \sum_{j=1}^{N} f(y_{ij})$ for $i = 1, \ldots, N$, where y_{ij} is the $[(i-1)N + j]$-th element of $\mathbf{y}$, and $f(\cdot)$ is a twice continuously differentiable function.[1]

[1] The problem can be generalized in a straightforward manner to linear combinations of the form $\mathbf{Y}_i = \sum_{j=1}^{N} \alpha_j f(y_{ij})$ for $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, N$. We consider the case of simple aggregation ($\alpha_j = 1$) for simplicity.

Considering the aggregation of non-linearly transformed bilateral variables (as would the case if we observe aggregated trade data at the country level and want to consider a bilateral gravity model of trade in log form), we can write the model for the aggregated variable as

$$\mathbf{Y} = \mathbf{A}(\mathbf{y}) = (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}), \tag{2}$$

where $\boldsymbol{\iota}_N$ is an N-dimensional column vector of ones and $\mathbf{f}(\cdot)$ is an N^2-dimensional vector function where $\mathbf{f}(\mathbf{y})$ with a typical element given by $f(y_{ij})$. Without loss of generality, we consider a square structure with a total of N^2 observations. In a gravity context, ruling out 'self-relationships' the number of observations would typically amount to $N(N-1)$.

We consider two alternative estimators of the parameters in equation (1) based on approximations of the nonlinear linkage between the bilateral and aggregate bilateral dependent variable. The first method relies on approximating the nonlinear aggregation relationship by ignoring the stochastic component of the bilateral relationship and using ML estimation. The second method relies on GLS estimation on a first-order Taylor expansion of equation (2).

2.1 Approximation I: Aggregate ML Estimation

A simple approach to estimating the parameters in equation (2) is based on an interpretation of the aggregated model as being affected by shocks at the aggregate instead of the bilateral level. The true model given by equation (2) can thus be thought of as being approximated by the specification

$$\mathbf{Y} = (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta}) + \boldsymbol{\eta}, \tag{3}$$

where $\boldsymbol{\eta} \sim N(\mathbf{0}, \sigma^2_\eta \mathbf{I}_N)$ is assumed. The normality assumption implies that the nonlinear least square estimator of $\boldsymbol{\beta}$ in equation (3),

$$\hat{\boldsymbol{\beta}}_{LS} = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\,(\mathbf{Y} - (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta}))'\,(\mathbf{Y} - (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta})) \tag{4}$$

is also the maximum likelihood estimator, $\hat{\boldsymbol{\beta}}_{ML}$, which is obtained as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{ML} = (\hat{\boldsymbol{\beta}}_{ML}, \hat{\sigma}^2_{\eta,ML}) = \underset{\boldsymbol{\beta},\sigma^2_\eta}{\operatorname{argmax}}\; L(\boldsymbol{\beta}, \sigma^2_\eta) = \\ = \underset{\beta,\sigma^2_\eta}{\operatorname{argmax}}\; \frac{1}{(2\pi\sigma^2_\eta)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2_\eta}\,(\mathbf{Y} - (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta}))'\,(\mathbf{Y} - (\mathbf{I}_N \otimes \boldsymbol{\iota}_N')\,\mathbf{f}(\mathbf{X}\boldsymbol{\beta}))\right\}\end{aligned} \tag{5}$$

The maximum likelihood estimator of the (aggregate) error variance can be obtained in a straightforward manner by setting the derivative of the log-likelihood with respect to σ_η^2 to zero, thus resulting in

$$\hat{\sigma}^2_{\eta,ML}(\boldsymbol{\beta}) = \frac{1}{N}\left(\mathbf{Y} - \left(\mathbf{I}_N \otimes \boldsymbol{\iota}'_N\right)\mathbf{f}(\mathbf{X}\boldsymbol{\beta})\right)'\left(\mathbf{Y} - \left(\mathbf{I}_N \otimes \boldsymbol{\iota}'_N\right)\mathbf{f}(\mathbf{X}\boldsymbol{\beta})\right), \tag{6}$$

which can be used to concentrate the log-likelihood with respect to σ_η^2 and obtain the maximum likelihood estimator of $\boldsymbol{\beta}$ as a solution to

$$\hat{\boldsymbol{\beta}}_{ML} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\; -\frac{N}{2}\left(\mathbf{Y} - \left(\mathbf{I}_N \otimes \boldsymbol{\iota}'_N\right)\mathbf{f}(\mathbf{X}\boldsymbol{\beta})\right)'\left(\mathbf{Y} - \left(\mathbf{I}_N \otimes \boldsymbol{\iota}'_N\right)\mathbf{f}(\mathbf{X}\boldsymbol{\beta})\right). \tag{7}$$

The solutions given by equations (4) and (7) are equivalent and the optimization problems can be solved using standard algorithms.

2.2 Approximation II: Linearized GLS

The setting given by equation (2) corresponds to the case of models of non-linearly aggregated data which can be nested within the class of models investigated by Proietti (2006). An estimate of for $\boldsymbol{\beta}$ can be obtained using a linearized version of (2). In particular, the Taylor expansion of equation (2) around some value of $\mathbf{y}$, $\bar{\mathbf{y}}$, is given by

$$\mathbf{Y} \approx \bar{\mathbf{Y}} + \boldsymbol{\Theta}(\bar{\mathbf{y}})(\mathbf{y} - \bar{\mathbf{y}}), \tag{8}$$

where $\boldsymbol{\Theta}(\mathbf{x})$ is the $N \times N^2$ Jacobian matrix of $\mathbf{A}(\mathbf{x})$. Proietti (2006) proposes an iterative estimation method for $\boldsymbol{\beta}$. Starting with a trial value of $\tilde{\mathbf{y}}$, the vector $\boldsymbol{\beta}$ can be estimated using

$$\hat{\boldsymbol{\beta}} = \left[(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\mathbf{X})'\left(\boldsymbol{\Theta}(\tilde{\mathbf{y}})'\boldsymbol{\Theta}(\tilde{\mathbf{y}})\right)^{-1}(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\mathbf{X})\right]^{-1}(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\mathbf{X})'\left(\boldsymbol{\Theta}(\tilde{\mathbf{y}})'\boldsymbol{\Theta}(\tilde{\mathbf{y}})\right)^{-1}\left(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\tilde{\mathbf{y}} + \bar{\mathbf{Y}} - \mathbf{Y}\right) \tag{9}$$

and the residuals at the bilateral level are given by

$$\hat{\boldsymbol{\varepsilon}} = \boldsymbol{\Theta}(\tilde{\mathbf{y}})'\left(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\boldsymbol{\Theta}(\tilde{\mathbf{y}})\right)^{-1}\left(\boldsymbol{\Theta}(\tilde{\mathbf{y}})\tilde{\mathbf{y}} + \bar{\mathbf{Y}} - \mathbf{Y} - \boldsymbol{\Theta}(\tilde{\mathbf{y}})\mathbf{X}\hat{\boldsymbol{\beta}}\right). \tag{10}$$

Subsequently, the variance of the error term can be estimated as $\hat{\sigma}^2 = \hat{\boldsymbol{\varepsilon}}'\hat{\boldsymbol{\varepsilon}}/N^2$. The fitted values of the unobserved bilateral variable $\hat{\mathbf{y}} = \tilde{\mathbf{X}}\hat{\boldsymbol{\beta}} + \hat{\boldsymbol{\varepsilon}}$ are then used as the next trial value and the procedure is repeated until the change in the fitted bilateral variable is sufficiently small.

It should be noted that the bilateral setting considered here differs from the usual time series applications in one important respect. In the case of decomposing, for instance, annual into quarterly data and having 20 years of observations, the decomposition ratio would be 4 over 20 (assuming that the same seasonal pattern holds for each year). In a cross-sectional, bilateral setting where each of the N aggregate observations has to be decomposed into $N-1$ bilateral observations, the decomposition ratio is approximately (and asymptotically exactly) equal to one. The simulation results presented in the following section can thus be be seen as a check of the ability of linearized GLS estimation in this more demanding setting.

3 Monte Carlo Results

We asses the performance of the two methods presented in the previous section using a simple Monte Carlo simulation exercise. We start by creating bilateral data using the following data generating process,

$$y_{ij} = 2 + 1.5x_{1,ij} + 1x_{2,ij} + 0.5x_{3,ij} + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim \text{NID}(0,1), \tag{11}$$

where $x_{1,ij}, x_{2,ij}$, and $x_{3,ij}$ are drawn from a standard normal distribution (treated as fixed in repeated samples), making the signal-to-noise ratio amount to 3.5. The bilateral dimension of the data ranges from $i = 1, \ldots, N$ and $j = 1, \ldots, J$, yielding a total IJ observations. Two sample sizes will be considered: $I = 20$ and $J = 19$ (280 observations), as well as $I = 50$ and $J = 49$ (2450 observations).

The simulated bilateral data are transformed and summed up into I aggregate observations, assuming that model (11) is specified in log form, i.e., $Y_i = \sum_{j=1}^{J} \exp(y_{ij})\ \forall i = 1, \ldots, I$ (setup 1). Alternatively, we add a (normal) error term (τ_i) to the aggregate data (Y_i), reflecting a possible mismatch between the true bilateral and the aggregate data, e.g., due to missing observations that have to be imputed (setup 2). The error in the aggregate data is assumed to have a standard error equal to one fifth of that of the aggregate data.

Having generated the bilateral and aggregate data on the dependent variable, we compare three estimation methods: i) a standard least squares regression of the bilateral data (y_{ij}) on the bilateral explanatory variables (Bilateral LS, BLS), which serves as a benchmark; ii) the ML estimates based on the aggregate data of the dependent variable and the disaggregated information of the explanatory variables as outlined in section 2.1 (Aggregate ML, AML); and iii) combining the aggregate data on the dependent variable (Y_i) with the bilateral data on the explanatory variables according to the approach outlined in section 2.2 (Linearized GLS, LGLS).

Table 1 shows the average bias and RMSE of the three estimation methods based on 1,000 replications and reports the correlation between the (levels of the) actual and predicted aggregate values (ρ_A), as well as the correlation between the (levels of the) actual and predicted bilateral values (ρ_B). Since the simulation results for the three slope parameters are very similar, we report only their average bias and RMSE, along with those of the constant for the sake of brevity.

Table 1: Monte Carlo results: Estimates based on bilateral versus aggregated data

		BLS		AML		LGLS	
Setup 1		Bias	RMSE	Bias	RMSE	Bias	RMSE
20×19	Constant	0.004	0.053	3.442	3.467	-0.113	0.446
	Slope parameters	0.001	0.054	-0.144	0.441	0.078	0.476
	ρ_A	0.581		0.673		0.999	
	ρ_B	0.681		0.654		0.777	
50×49	const.	-0.001	0.020	4.465	4.475	0.056	0.184
	x (av.)	-0.001	0.020	-0.130	0.400	-0.006	0.522
	ρ_A	0.677		0.687		0.998	
	ρ_B	0.685		0.659		0.824	
Setup 2							
20×19	Constant	0.002	0.052	3.325	3.339	-0.016	0.349
	Slope parameters	-0.001	0.054	-0.078	0.381	0.077	0.465
	ρ_A	0.708		0.777		0.997	
	ρ_B	0.709		0.708		0.768	
50×49	const.	0.000	0.021	4.379	4.384	0.041	0.209
	x (av.)	0.000	0.021	-0.043	0.220	0.074	0.287
	ρ_A	0.691		0.718		0.999	
	ρ_B	0.662		0.667		0.787	

Notes: Simulation results based on 1,000 replications. See text for a detailed description.
BLS: OLS based on bilateral data.
AML: ML based on aggregated data.
LGLS: linearized GLS based on aggregated data.
ρ_A (ρ_B) … correlation between the actual and predicted aggregate (bilateral) values.

The results for both setups indicate that, as expected, the direct estimator, which is always the preferred choice if bilateral data are available, performs best in terms of the bias and RMSE, which essentially fade away for the large sample considered. When comparing the aggregate ML and the linearized GLS approach, the latter is clearly superior in estimating the intercept term. Given that the identification of the intercept term through the AML estimator is exclusively based on functional form, this result is not particularly surprising.

Regarding the slope parameters, the two methods perform equally well. With an average slope parameter of 1, the average magnitude of the bias ranges from 0.6% to 14%. This suggest that the approximation procedures based on aggregate data on the dependent variable provide at least a reasonable indication of the magnitude of the bilateral model parameters.[2]

The linear GLS estimator, apparently as a result of imposing the summing up constraint, stands out in generating by far the highest correlation between the (level of the) actual and predicted values, both for the bilateral and aggregate values. Hence, for the purpose of decomposing aggregate values and generating unobserved bilateral data that can be used in further regression analyses, the linearized GLS estimator would be the recommended choice. An interesting application of the method would be related to generating bilateral linkage (weight) matrices in spatial econometric studies. In spatial econometric models, the matrix of spatial linkages is often row-normalized, a transformation that can be expected to mitigate the error in the predicted values and a high correlation of the predicted values with the actual (unobserved) elements of the weights matrix could be seen as the most important goal in order to produce reasonably good approximations.

4 Application: Intra-EU Trade and Capital Flows

In this section we provide a small scale empirical application of the methods for estimating models based on bilateral relationships using aggregate data. We employ cross-sectional data for EU15 countries. The use of a set of highly integrated, developed countries justifies to some extent the use of the simplest gravity model, using as explanatory variables of the bilateral trade variable the distance between the country of origin (i) and the destination country (j) and their combined size. Hence, the bilateral model considered is given by

$$y_{ij} = \beta_0 + \beta_1 \ln GDP_{ij} + \beta_2 \log DIST_{ij} + \varepsilon_{ij}, \tag{12}$$

where $\ln GDP_{ij} \equiv \ln GDP_i + \ln GDP_j$ is the sum of the (log of the) two countries' GDP. As dependent variable y_{ij}, we consider several alternatives. First, we use the log of exports from country i to country j (EX_{ij}); this variable is observed both at the bilateral and aggregate level, such that the direct an indirect estimates can be compared. Second, we use measures of financial openness, derived from the capital account, which are not available at the bilateral but only at the aggregate level. In particular, we consider (i) the log of total asset claims of country i against country j (TA_{ij}), (ii) the log of the stock of foreign direct investment of country i in country j (FDI_{ij}), and iii) the log of portfolio investment of country i in country j (PI_{ij}).

Our cross-section dimension comprises 14×13 countries[3], yielding a total of 182 observations evaluated in the year 2005. Data on bilateral exports, distance and GDP are from the CEPII

[2] We also considered the size of standard t-test. Both approximation methods show severe distortions in terms of the estimation of the variance of the estimate, suggesting that they cannot be reasonably used for inference on the parameters in the bilateral model beyond obtaining a point estimate.

[3] Belgium and Luxembourg are treated as a single economy for reasons of data availability.

gravity dataset.[4]. Aggregate data used for the indirect estimation are taken from the World Bank's WDI database (EX) and from Lane and Milesi-Ferretti (2007) (TA, FDI, PI).

Table 2: Estimation results: trade, assets, FDI and portfolio investment (EU15, 2005)

	EX			TA		FDI		PI	
	BLS	AML	LGLS	AML	LGLS	AML	LGLS	AML	LGLS
const.	-2.715	-0.679	-4.966	0.538	-4.911	-3.625	-7.539	0.555	-5.64
$\ln GDP_{ij}$	0.788	0.794	0.885	0.843	0.949	0.963	1.072	0.782	0.912
$\ln DIST_{ij}$	-1.366	-1.255	-1.375	-1.331	-1.351	-1.451	-1.717	-1.217	-1.223
ρ_A	0.926	0.929	0.999	0.854	0.999	0.869	0.999	0.876	0.999
ρ_B	0.900	0.899	0.887	-	-	-	-	-	-

Notes: Estimates based on $I = 14$ and $J = 13$ (182) observations.
EX ... Exports, TA ... Total Assets , FDI ... Foreign Direct Investment, PI ... Portfolio Investment.
BLS ... Bilateral Least Squares, AML ... Aggregate ML, LGLS ... Linearized GLS.
ρ_A (ρ_B) ... correlation between the actual and predicted aggregate (bilateral) values.

Table 2 gives an overview of the estimation results. The first three columns report the results of the BLS, AML, LGLS) using exports as dependent variable. The bilateral least squares estimates confirm our expectation: the size variable (GDP_{ij}) enters with a positive coefficient close to unity, whereas distance ($DIST_ij$) has a strong negative effect on bilateral exports with an elasticity of -1.36. Both the AML and the LGLS estimates, derived from aggregate export data (and bilateral data on the explanatory variables) replicate these results quite well in the sense that their point estimates for the parameters of GDP_{ij} and $DIST_{ij}$ are quite close to those obtained with bilateral data. The LGLS estimate is closer to the BLS estimate of the parameter of $DIST_{ij}$, whereas the AML estimate is closer the the BLS estimate of the parameter of GDP_{ij}. Hence, in terms of the possible bias, there appears to be no clear preference for one of the estimators, as was already inferred from the results of the Monte Carlo study presented in section 3.

For the present application, the performance of the two methods in generating the bilateral data is virtually identical. The bilateral data implied by the AML and LGLS estimates of the model with aggregated data are highly correlated with the actual bilateral data; in fact the correlation of 0.89 is practically identical to that implied by the BLS estimation. Figure 1 compares the true bilateral data with the predicted values from the LGLS estimation. The fit of the data is quite precise and there is no evidence of a systematic under- or overestimation of trade flows over large subsets of the sample.

Turning to the results for the AML and LGLS estimates for the financial openness measures, for which bilateral data do not exist, the point estimates of the parameters appear intuitively appealing. The coefficient on the composite GDP variable remains close to unity, whereas the coefficient of the distance becomes larger in magnitude for foreign direct investment and smaller in magnitude for portfolio investment as compared to the results for the trade

[4] `http://www.cepii.fr/CEPII/en/bdd_modele/presentation.asp?id=8`

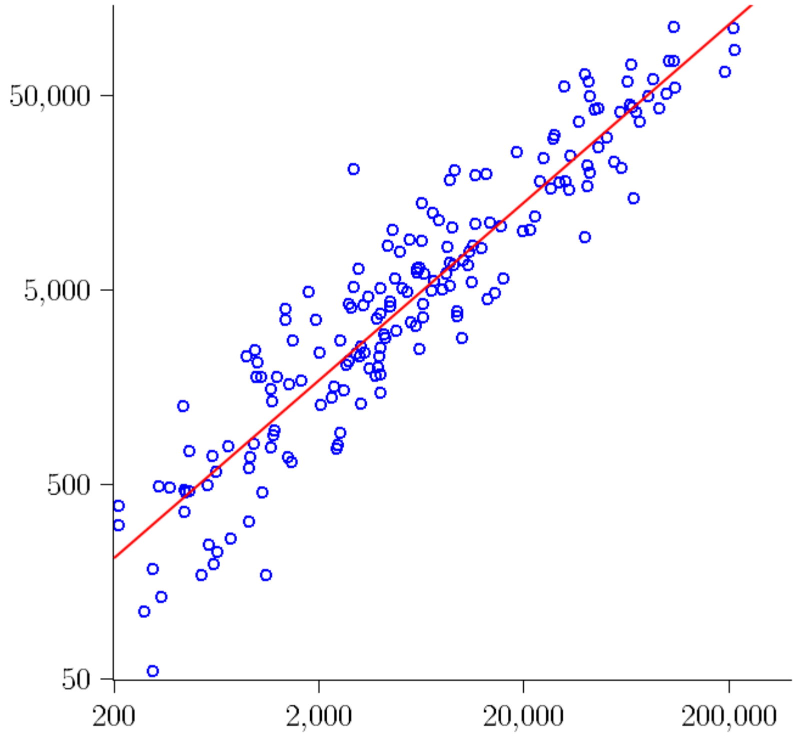

Figure 1: Actual values (horizontal axis) versus LGLS fitted bilateral values (log scale)

variable. This suggests that FDI, involving a larger engagement than portfolio investment (both in magnitude and the intention to exert influence on the business operation) is more affected by a larger distance, which can thus be interpreted not only as a proxy for trade costs, but also a proxy for differences in culture, legal systems and difficulties in enforcing property rights.

For total assets, being comprised of both FDI and PI, these changes apparently offset each other, resulting in a coefficient of $DIST_{ij}$ that is very close to the one for the export equation. When considered in the context of existing empirical gravity models for financial flows, these results are plausible (see e.g. Portes et al., 2001, for international transactions in financial assets) and underline the applicability of the two methods put forward.

General statements about the performance of the indirect estimators are difficult. Their properties will depend on the sample size, the complexity of the data generating process, the number of explanatory variables considered, the nonlinearities involved in the aggregation, the presence of outliers, and the properties of the error term. However, in sum both the Monte Carlo results and the application considered in this paper suggest that the approximation methods can work reasonably well and should be considered as a first step towards empirically investigating models of policy interest where bilateral data is not available.

5 Conclusions

This paper considers alternative methods to estimate bilateral models when only aggregate data on the dependent variable is available. The purpose of such an indirect approach is twofold. First, it can be used to obtain indicative results on bilateral model parameters. Second, the methods can be utilized to decompose aggregate into bilateral data, based on an established empirical model. We show, using both a Monte Carlo study and a simple application to intra-EU trade and capital flows, that the indirect estimation methods work reasonably well and are worthwhile being considered further in empirical research.

The methods proposed have been tested in the simulation under the assumption of a known data generating process and thus abstract from potential biases arising from misspecification. Both the results of the Monte Carlo simulation and the empirical application indicate that the application of the methods proposed can be helpful to obtain estimates of disaggregated data if (a) the nonlinear aggregation procedure is known and (b) if there exists a group of predictors of the disaggregated variable which is well known to explain its variation well. In this respect, international bilateral flows of goods and persons, where gravity models are well known to provide a good fit to the data, appear as suitable candidates for applications (see Crespo Cuaresma et al., 2013, for a recent application of similar methods to migration flow estimation).

Several potentially fruitful directions for future research can be highlighted. Alternative estimation methods and more comprehensive models should be investigated. In addition, a question of interest relates to the asymptotic properties of the indirect estimators and the assumptions required for asymptotic equivalence of the direct and indirect estimators. Finally, since it is often the case that a subset of bilateral data is available, modifications of the estimation methods considered here that are able to exploit this additional information would be of interest.

References

Abel, G. (2013). Estimating Global Migration Flow Tables Using Place of Birth Data. *Demographic Research*, 28:505–546.

Crespo Cuaresma, J., Moser, M., and Raggl, A. (2013). On the Determinants of Global Bilateral Migration Flows. WWWforEurope Working Papers Series 5.

Frankel, J. A. and Romer, D. (1999). Does Trade Cause Growth? *American Economic Review*, 89:379–399.

Granger, C. and Lee, T.-H. (1999). The effect of aggregation on nonlinearity. *Econometric Reviews*, 18(3):259–269.

Lane, P. R. and Milesi-Ferretti, G. M. (2007). The external wealth of nations mark ii: Revised and extended estimates of foreign assets and liabilities, 1970-2004. *Journal of International Economics*, 73(2):223–250.

Lee, T.-H., White, H., and Granger, C. W. J. (1993). Testing for neglected nonlinearity in time series models : A comparison of neural network methods and alternative tests. *Journal of Econometrics*, 56(3):269–290.

Mitze, T., Alecke, B., and Untiedt, G. (2010). Trade-fdi linkages in a simultaneous equations system of gravity models for german regional data. *International Economics*, 122:121 – 162.

Portes, R., Rey, H., and Oh, Y. (2001). Information and capital flows: The determinants of transactions in financial assets. *European Economic Review*, 45(4-6):783 – 796.

Proietti, T. (2006). On the Estimation of Nonlinearly Aggregated Mixed Models. *Journal of Computational and Graphical Statistics*, 15:18–38.

Sarisoy Guerin, S. (2006). The Role of Geography in Financial and Economic Integration: A Comparative Analysis of Foreign Direct Investment, Trade and Portfolio Investment Flows. *World Economy*, 29:189–209.

Silvestrini, A. and Veredas, D. (2008). Temporal Aggregation Of Univariate And Multivariate Time Series Models: A Survey. *Journal of Economic Surveys*, 22(3):458–497.