Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics and Operations Research

Clustering Exchangeable Relational Arrays

V.A.D. Goorden (548046)

Supervisor:	P. Wan
Second assessor:	A. Archimbaud
Date final version:	9th July 2023

Abstract

Exchangeability within relational arrays signifies the fundamental property that the statistical distribution of the relations is not affected by a reordering of the data. The exchangeable attribute is often an intrinsic feature of relational data, which in turn systematically presents interactions of pair-wise relationships. These arrays allow applicability in versatile applications across diverse domains. This research builds upon prior work on exchangeable relational arrays. Emphasis is placed on estimating the covariance structure in complex relationships that accommodate dependencies among entities or within specific groups. The exchangeability is exploited by performing both an intuitive and empirical clustering analysis to enhance predictability by capturing the dynamics within clusters. Findings reveal that the clustered exchangeable estimator enhances inference through interpretability and model fit. Thereupon, the scope of the exchangeable estimator is extended to more heterogeneous data.

Key words: Clustering; Relational data; GEE; International trade

1 Introduction

Relational arrays are characterised by pair-wise interactions between entities and are quantified regarding the type of the relation. Relational arrays can be modeled by a binary relation, an ordinal relation, which captures the strength of such an interaction, or a continuous relation, which provides a spectrum of potential values. Hence, these can be used for a large scope of instances.

In many applications, relational data occurs, e.g. transport networks (Chen et al., 2020), marketing (Zahay et al., 2004), biological systems (Walsh et al., 2020), financial (Dicken et al., 2001) and social networks (Attanasio et al., 2012). Using relational arrays has the advantage of providing a structured representation of existing relations between entities. Applying statistical and machine learning methods facilitates analysis and prediction, which necessitates the formulation of models that effectively capture intricate relationships by estimating coefficients. However, the core challenge for the estimation purpose lies in taking into account the interdependence that arises among relations containing shared actors or belonging to specific groups. Accommodating these interdependencies in the covariance structure poses another difficulty. Moreover, precise estimation of standard errors calls for additional assumptions to alleviate the complexity (Fosdick & Hoff, 2014).

In our research, we leverage an exchangeability assumption which is often inherent in relational arrays. Concisely, exchangeability refers to the property that the ordering of the observations does not influence the distribution of the relationships. This assumption has been prevalent in the analysis of dyadic data such as Lloyd et al. (2013), Crane & Dempsey (2019), and Fan et al. (2020). However, Marrs et al. (2023) only first used the assumption as a tool to adjust for the dependencies. Due to this incorporation, the estimation of the covariance structure becomes more homogeneous.

This paper commences with a replicative study on Marrs et al. (2023). The first part entails a simulation study using multiple data generation techniques to compare confidence intervals. The estimators considered are one that assumes exchangeability and one that does not, which is the dyadic clustering estimator (Fafchamps & Gubert, 2007; Aronow et al., 2015). The second part incorporates international trade data by Westveld & Hoff (2011) in an out-of-sample prediction study using ordinary least squares (OLS) and the exchangeable estimator.

To extend to the aforementioned paper, we aim to investigate how to account for heterogeneity within an exchangeable framework. Hence, we dissect the data by means of $k-$means clustering such that across clusters, the observations can not be permuted to maintain the same probability density. However, within a cluster, there exists exchangeability. This can be understood from the perspective of a social network in which diverse communities exist. Here, the interactions in a certain group can have a different covariance structure than in another group, yet within this particular group, there is homogeneity. The k clusters that follow each form an exchangeable covariance estimator similar to Marrs et al. (2023).

This leads to the following research question: *"How can we effectively model and analyse exchangeable relational arrays that contain heterogeneity to capture the distinct relational patterns exhibited by different subgroups?"*. Finally, the clustered exchangeable estimator is compared to the exchangeable estimator in terms of prediction error and model fit using the international

trade data.

The conclusions that follow from this tripartite research are that: (1) The exchangeable estimator has on average a much smaller bias compared to the dyadic clustering estimator in both an exchangeable and non-exchangeable data setting. (2) When analysing real-world data, the exchangeable estimator yields a much higher R^2 than a simple linear regression. And, (3), clustering the data has additional value to improve the fit of the model in comparison to non-clustered analysis. The best way to incorporate information provided by the clusters proved to be adding binary variables of the clusters, which resulted in a large increase in R^2.

Our contribution to the current literature comprises extending the scope of the exchangeable estimator of Marrs et al. (2023) for a wider variety of data. The clustered exchangeable estimator addresses differences in the data but maintains the desirable properties. Above that, it enhances interpretability due to the clusters that have distinct covariances. While an increasing amount of literature is devoted to modelling exchangeable relational arrays, fewer models are available for heterogeneous relational arrays (Fosdick, 2013). Therefore, we combine the two to ensure real-world applicability. In addition, advanced models enhance prediction and inference on future relationships, which in turn improves decision-making.

This paper proceeds in the following manner: First, the previous literature and the topic of exchangeable relational arrays will be discussed in more depth in Section 2. Then, we will touch upon the data used in Section 3. Furthermore, the methods are described in Section 4 after which the results are mentioned in Section 5. Consequently, we conclude in Section 6.

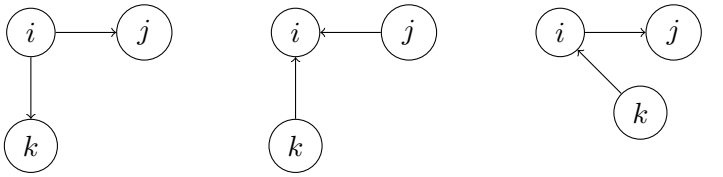
2 Theory

First, notation is introduced. A relational array is given by $Y = [y_{ijr}]$, where i denotes the sending actor, and j the receiving actor in a directed relation for $i, j = 1, ..., n$ with $i \neq j$. $r = 1, ..., R$ resembles the relational context r, or the time period, which in that case can also be denoted by $t = 1, ..., T$. The relational array Y is composed of all the R matrices. These matrices are of size $(n \times n)$ and each describes the paired relations in each context r out of the n actors. Since the relations are directed, it holds that, in general, for a given r, $y_{ij} \neq y_{ji}$. Therefore, the matrix of Y is not symmetrical. However, there still exists a symmetric covariance matrix structure.

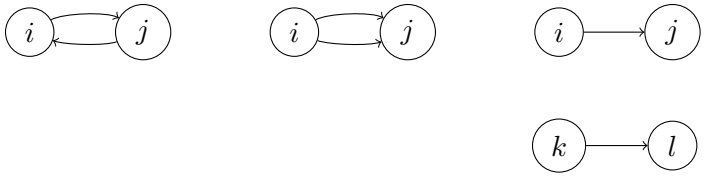
The dependencies within relational data influence the parameters in the covariance matrix. Beck et al. (2006) analysed these possible dependencies in a trade setting and found that there exist more dependencies than just reverse dyad reliance. In Figure 1, the six potential pairs are depicted for a distinct set of actors i, j, k, l represented by nodes. Each arrow embodies a directed array. In each sub-figure, the arrows can either be corresponding to the same relational context $r = s$, or two different contexts $r \neq s$. From this follows a set of six distinct parameters when $r = s$ and six when $r \neq s$. Therefore, there are twelve different parameters in the covariance matrix Ω.

First, (a) describes the dependence that exists when two relations share a common sender. For instance, the dependence when Greece exports to the US and to the UK in the same period. In this time period, Greece could for instance occur in a recession, and, therefore, there is an interdependence in all arrays that originate from Greece. Additionally, the size of the flow is

Figure 1: *The six forms of dependencies between relational pairs*



(a) Same sender (b) Same receiver (c) Common actor



(d) Reciprocity (e) Observational dependence (f) No dependence

Note. both directed arrows can either be corresponding to the same relational context $r = s$, or to different contexts $r \neq s$.

also dependent on the originating country, which maintains across different time periods. By a similar argument, the dependence in (b) can be understood. Put differently, some countries consume more than other countries. In (c) the situation with a shared actor is depicted. This could differ across countries that have more open economies versus closed economies. Moreover, (d) captures a twofold relationship. In case both the sender and the receiver are in a trade agreement, this would result in a different parameter as compared to two nations that are in conflict. Another form of dependence is given in (e) which mainly occurs when the two directed arrows corresponding to r and s are unequal; $r \neq s$. In this case, we encounter temporal dependence of the directed array. Lastly, (f) is included for completeness, which represents the independent case since there are no shared actors.

Furthermore, these twelve parameters affect the covariance matrix. Since we assume that two non-overlapping pairs are independent, the parameters belonging to case (f) are set to zero. We denote the set of parameters by $\phi_u^{(\eta)}$, where u is equal to one of the six dependency cases a, b, c, d, e, f, with $\phi_f = 0$. η indicates whether the relational context is the same or not; $\eta = 1$ for $r = s$, and $\eta = 2$ for $r \neq s$. In Table 1, the covariance structure is depicted for a single relational context. These entries include the five non-zero parameters for $r = s$. The matrix can be enlarged for the case when $r \neq s$ with the parameters $\phi_{\eta=2}$.

2.1 Literature Review

Warner et al. (1979) wrote the seminal paper in this field of research and proposed a method of moments procedure to estimate dyadic data. Hoff (2003) introduced a generalized linear model for network data to deal with the three forms of data structures: binary, ordinal or continuous. Additionally, an important breakthrough for relational modelling was the notion of statistical dependencies. Often, the interconnected entities are treated as independent occurrences, however, Hoff & Ward (2004) illustrates that there might be a correlation with relationships with

Table 1: *The six different dependency parameters in the exchangeable covariance matrix* $\Omega_E =$

	Y_{AB}	Y_{AC}	Y_{AD}	Y_{BA}	Y_{BC}	Y_{BD}	Y_{CA}	Y_{CB}	Y_{CD}	Y_{DA}	Y_{DB}	Y_{DC}
Y_{AB}	σ^2	ϕ_a	ϕ_a	ϕ_d	ϕ_c	ϕ_c	ϕ_c	ϕ_b	0	ϕ_c	ϕ_b	0
Y_{AC}	ϕ_a	σ^2	ϕ_a	ϕ_c	ϕ_b	0	ϕ_d	ϕ_c	ϕ_c	ϕ_c	0	ϕ_b
Y_{AD}	ϕ_a	ϕ_a	σ^2	ϕ_c	0	ϕ_b	ϕ_c	0	ϕ_b	ϕ_d	ϕ_c	ϕ_c
Y_{BA}	ϕ_d	ϕ_c	ϕ_c	σ^2	ϕ_a	ϕ_a	ϕ_b	ϕ_c	0	ϕ_b	ϕ_c	0
Y_{BC}	ϕ_c	ϕ_b	0	ϕ_a	σ^2	ϕ_a	ϕ_c	ϕ_d	ϕ_c	0	ϕ_c	ϕ_b
Y_{BD}	ϕ_c	0	ϕ_b	ϕ_a	ϕ_a	σ^2	0	ϕ_c	ϕ_b	ϕ_c	ϕ_d	ϕ_c
Y_{CA}	ϕ_c	ϕ_d	ϕ_c	ϕ_b	ϕ_c	0	σ^2	ϕ_a	ϕ_a	ϕ_b	0	ϕ_c
Y_{CB}	ϕ_b	ϕ_c	0	ϕ_c	ϕ_d	ϕ_c	ϕ_a	σ^2	ϕ_a	0	ϕ_b	ϕ_c
Y_{CD}	0	ϕ_c	ϕ_b	0	ϕ_c	ϕ_b	ϕ_a	ϕ_a	σ^2	ϕ_c	ϕ_c	ϕ_d
Y_{DA}	ϕ_c	ϕ_c	ϕ_d	ϕ_b	0	ϕ_c	ϕ_b	0	ϕ_c	σ^2	ϕ_a	ϕ_a
Y_{DB}	ϕ_b	0	ϕ_c	ϕ_c	ϕ_c	ϕ_d	0	ϕ_b	ϕ_c	ϕ_a	σ^2	ϕ_a
Y_{DC}	0	ϕ_b	ϕ_c	0	ϕ_b	ϕ_c	ϕ_c	ϕ_c	ϕ_d	ϕ_a	ϕ_a	σ^2

Note. this structure represents the case for $R = 1$. The matrix is symmetric and contains a zero for every pair of independent relationships.

the same actor in dyadic data that should be taken into account. Yet, the difficulty lies in how to model these correlated error terms which capture the dependencies.

One approach is to impose a parametric latent variable structure on the errors (Hoff, 2005). The downfall of this approach is that the specified parametric model is not always consistent with true error structure (Fosdick & Hoff, 2014). Additionally, it is computationally heavy to calculate these models since it often involves Markov Chain Monte Carlo.

The second, and more model-agnostic approach, accounts for the relational dependence by empirically estimating the error structure based on residuals from the regression. This idea originates from Conley (1999) and Fafchamps & Gubert (2007) further implemented this. In Aronow et al. (2015), the dyadic clustering estimator was finally proposed as a sandwich-type variance estimator. This estimator is non-parametric and should account for complex cluster structures that are present within relational data. Hereafter, Carlson et al. (2021) published an article using the dyadic clustering estimator to investigate international relations and find that dyadic clustering leads to an underestimation of uncertainty. Moreover, this estimator rests on multiple conditions to acquire consistency. On top of that, there still exist limitations in this approach due to high variability in the standard errors.

Subsequently, we define a new approach to obtain more robust standard errors based on an assumption that is fundamental to many relational datasets. Namely, we impose an exchangeability assumption. This property, which was first introduced by De Finetti, is desirable in modelling since it simplifies the analysis and allows the use of more general models.

Exchangeability has been extended by Hoover and Aldous (1981) for relational arrays. The idea is similar since it must hold that the distribution of the data remains the same after any permutation or rearrangement of the rows and columns. The difference lies in that for non-relational data, it is about the ordering and shuffling of individual observations, whereas for relational data, the shuffling should only be applied to the arrangement of entities or a simultaneous reordering of the third dimension. Formally, the probability distribution of the

error array $p(\mathcal{E}) = p(\Pi(\mathcal{E}))$, where $\Pi(\mathcal{E}) = (\epsilon_{\pi(i)\pi(j)\nu(r)})$ indicates the reordered array of errors according to permutation operators π and ν, should remain unchanged. Note that data with time as the relational context can still be exchangeable if the majority of the temporal variability is explained through the covariates.

The assumption of exchangeability has widely been used in forecasting relational arrays, with applications in machine learning and Bayesian models, (Hoff, 2009; Lloyd et al., 2013; Cai et al., 2016).Yet, incorporation of this property in the covariance structure was only first introduced by Marrs et al. (2023).

2.2 Clustered Exchangeability

Another field of research explored heterogeneous relational arrays and tackled the difficulty in predicting these by forming clusters. Long et al. (2007) provides a probabilistic framework to identify interaction patterns, among which one of the models is a k-means clustering algorithm. In the book of Long et al. (2010) a wider variety of these types of methods are described.

The novelty of this research is combining the exchangeable covariance feature with clustering analysis. It is essential to maintain exchangeability to keep the computation simple and efficient. These clusters are also to be incorporated into the exchangeable covariance matrix in Table 1. The key of this matrix stays the same, as interactions within a specified subgroup persist. Across clusters, when an actor from cluster A interacts with an actor in cluster B, new parameters should be designated. In this case, one can now interpret the directed array Y_{AB} as a cross-cluster interaction, with its own parameter. So in total, the covariance structure can be divided into a part with only interactions within a cluster, and a part with interactions outside of its cluster, in a similar manner as the extension of multiple relational contexts.

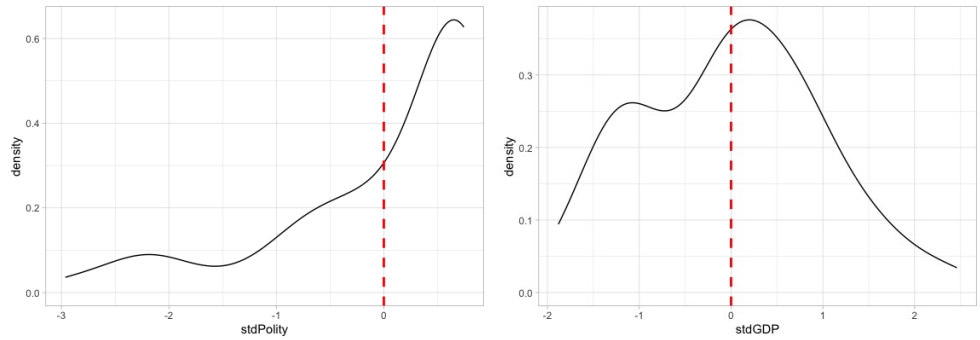
3 Data

After having described the structures that are present within relational data, we introduce the data we use, which is derived from Westveld & Hoff (2011). This dataset contains information on international trade between 58 countries over the period from 1981-2000, hence $R = T = 20$, which adds up to 66,120 observations. The dependent variable y_{ij} in this analysis is the log volume of trade from exporting country i to importing country j, such that $i \neq j$.

These data include six covariates and an intercept. The first two covariates represent the log of Gross Domestic Product (GDP) for both the exporting and importing countries. The log of the distance between the two countries is included and is inherently a constant. Additionally, two variables for the polity of a nation on both sides of the relation are added. These variables represent an ordinal relation with a measure ranging from 0, indicating a highly authoritarian, to 20, indicating a highly democratic state, and can be time-varying. Lastly, the variable cooperation in conflict captures whether, during a dispute, the nations were either on the same side, coded by $+1$, or vice versa coded by -1. When no dispute has taken place, or the parties did not take an active position in this, the value will be equal to zero.

Furthermore, this dataset is used as a starting point for the clustered exchangeable estimator. For the $k-$means clustering analysis we add an extra binary operator if the country is a European

Figure 2: *Distribution of the mean standardized polity and logGDP*



Union member state. Since most member states joined the EU between 1980-1995, we only take the data from 1995-2000 into account. This also results in more representative clusters due to a shorter, hence less fluctuating, time period to cluster over.

From the six covariates, we cluster on the logGDP and the polity of the exporting countries. The imported values are omitted since the clustering is based on the countries, not on the pairs. Therefore, also, the distance is left out, since this does not have an interpretation without a pair. However, we assume that the EU variable can account for distance partially. Additionally, cooperation in conflict is excluded since this variable is time-varying and we aim to find constant clusters. This is also the reason why the mean of the polity and the mean GDP over time are chosen.

Moreover, it is important to standardize the data when performing the cluster analysis due to two reasons: (1) If the variables have different scales, the clustering algorithm is influenced by the variables with larger scales. Since $k-$means clustering uses the Euclidean distance measure to calculate the similarity between data points, those larger-scaled variables can dominate the formation of clusters. Therefore, if standardization has been performed, all variables contribute equally to the clustering analysis. (2) Secondly, standardized data enhances the interpretability of clusters. Due to the similar scales, it is clear which observations are most influential and which groups exist within the data.

In Figure 2, the distributions of the mean standardized polity and log GDP are given. Clearly, polity is highly left-skewed, which influences the clustering analysis. The distribution of GDP looks relatively normal, which is confirmed by the Jarque-Bera test with a p-value of 0.604.

4 Methodology

The relational arrays are estimated by linear regression, given by:

$$y_{ijr} = \beta^T x_{ijr} + \epsilon_{ijr} \quad \text{for} \quad i, j = 1, ..., n, \ i \neq j, \ r = 1, ..., R, \tag{1}$$

where y_{ijr} measures the directed relation from i to j in context r, β contains the coefficients, x_{ijr} is a column vector of covariates of size $(p \times 1)$, and ϵ_{ijr} contains the unobserved errors.

The starting point to estimate the coefficients in the regression model is ordinary least squares

(OLS). For simplicity, the model is written in matrix notation:

$$Y = X\beta + \mathcal{E}, \tag{2}$$

where Y is the $R(n(n-1))$ vector of relations and X represents a matrix of size $(Rn(n-1) \times p)$. The OLS estimator is an unbiased estimator and has the smallest variance (BLUE) when there is independence across observations. However, since we deal with dependencies among similar actors, this is not the case. Therefore, we need to extend to the generalized least squares (GLS) estimator. The GLS estimator yields a BLUE estimator when the true value of Ω is known. Yet, this is in most practical instances not the case. Hence, we can account for this uncertainty in Ω by estimating the matrix $\hat{\Omega}$. This procedure is called feasible generalized least squares (FGLS) and proceeds as follows: first, OLS is performed to obtain the residuals $\hat{\epsilon}_{ijr} = (Y - X\hat{\beta}_{OLS})_{ijr}$, then Ω can be consistently estimated by the use of $\hat{\mathcal{E}}$. From this, the FGLS can be expressed as:

$$\hat{\beta}_{FGLS} = (X^T\hat{\Omega}^{-1}X)^{-1}X^T\hat{\Omega}^{-1}y, \tag{3}$$

with y being a vectorisation of (y_{ijr}). The estimate of the covariance matrix $\Omega = V(y|X)$ can then be constructed as:

$$\hat{V}_{FGLS} = (X^T\hat{\Omega}^{-1}X)^{-1}X^T\hat{\Omega}^{-1}\Omega\hat{\Omega}^{-1}X(X^T\hat{\Omega}^{-1}X)^{-1}, \tag{4}$$

this variance estimator is deduced as a sandwich estimator (Huber, 1967) due to the structure of two similar terms at the left and right side 'sandwiching' the FGLS and OLS estimator.

4.1 Dyadic clustering estimator

Fafchamps & Gubert (2007) and Aronow et al. (2015) formed the dyadic clustering estimator based on the single assumption that if two relations (i, j, r) and (k, l, r) do not share an actor, the relations are independent. Mathematically, this means $\text{cov}(y_{ijr}, y_{kls}|X) = \text{cov}(\epsilon_{ijr}, \epsilon_{kls}|X) = 0$ when there is no overlap in actors. From this relationship follows a covariance matrix with three types of inputs: (1) the diagonal elements represent the variance of relationship (i, j), (2) the off-diagonal covariance elements are zero whenever there is no common actor, and (3) no restrictions are placed when there is a common actor. The last property results in a variety of possible covariance values. The variance-covariance matrix based on this assumption is denoted by Ω_{DC}. To estimate the non-zero elements $\text{cov}(\epsilon_{ijr}, \epsilon_{kls})$ in Ω_{DC}, Fafchamps & Gubert (2007) suggested to use the product of residuals, namely $e_{ijr}e_{kls}$, where $e_{ijr} = y_{ijr} - \hat{\beta}^T x_{ijr}$, resulting in the covariance matrix $\hat{\Omega}_{DC}$. Sandwich variance estimation for $\text{var}(\hat{\beta}|X)$ is applied, which results in the dyadic clustering estimator $\hat{V}_{DC}$:

$$\hat{V}_{DC} = (X^TX)^{-1}X^T\hat{\Omega}_{DC}X(X^TX)^{-1}. \tag{5}$$

As mentioned before, the non-zero covariance elements are unrestricted and hence quite variable. This uncertainty comes from the estimation of $\hat{\Omega}_{DC}$, which has fewer observations than estimable elements.

4.2 Exchangeable estimator

Incorporating the exchangeability assumption in relational arrays, which is often inherent, yields the estimator for the covariance matrix that is derived in this section. First, we need to find estimates of the parameters $\phi_u^{(\eta)}$ for $u = a, b, c, d, e, f$ and $\eta = 1, 2$, which represent the covariances between relations. Since $\phi_f = 0$, only ten parameters in Ω need to be estimated. The parameters are approximated by taking an average of the products of the residuals with the same indexes. For instance, $\phi_a^{(2)}$, which is the covariance estimator of the same sender in two different contexts, can be specified as:

$$\phi_a^{(2)} = \binom{R}{2}^{-1} \frac{1}{n(n-1)(n-2)} \sum_{r \neq s} \sum_{i} \sum_{j \neq i} e_{ijr} (\sum_{k \neq j} e_{iks} - e_{ijs}). \tag{6}$$

The nine remaining estimators can be determined analogously.

Secondly, the estimator of the exchangeable covariance estimator $\hat{\Omega}_E$ is constructed in the following manner:

$$\hat{\Omega}_E = \sum_{\eta=1}^{2} \sum_{u=a}^{f} \hat{\phi}_u^{(\eta)} \mathcal{S}_u^{(\eta)}, \tag{7}$$

with $\mathcal{S}_u^{(\eta)}$ denoting a binary matrix of size $(Rn(n-1) \times Rn(n-1))$ which has the entry 1 for the relations that are of the corresponding type $u = a, b, c, d, e, f$ and $\eta = 1, 2$.

By similar reasoning, using the sandwich estimator, the estimator of $V(\hat{\beta}|X)$ under exchangeability is equal to:

$$\hat{V}_E = (X^T X)^{-1} X^T \hat{\Omega}_E X (X^T X)^{-1}. \tag{8}$$

The exchangeable estimator posits a moment-based feature, hence, it is consistent. In addition, the dyadic clustering estimator is highly parameterised and thus there is efficiency gain when using the exchangeable estimator in an exchangeable framework.

4.3 Simulation study

The two aforementioned estimators in Sections 4.1 and 4.2 are evaluated by means of a simulation study where data will be drawn from a linear regression model with either an exchangeable error model or a non-exchangeable error model. The regression model for both cases contains three covariates:

$$y_{ij} = \beta_1 + \beta_2 \mathbb{1}_{x_{2i} \in C} \mathbb{1}_{x_{2j} \in C} + \beta_3 |x_{3i} - x_{3j}| + \beta_4 x_{4ij} + \epsilon_{ij}, \tag{9}$$

with β_1 being the intercept. $\mathbb{1}$ denotes an indicator function which returns one if both i and j belong to a pre-specified subgroup C, and zero otherwise. β_2 corresponds to the coefficient belonging to the binary class-specific covariate. β_3 and β_4 respectively represent the coefficients for the positive real-valued actor-specific and the real-valued pair-specific covariates. The covariate x_{2i} is independently simulated from a Bernoulli(1/2) distribution, whereas x_{3i} and x_{4ij} are independently drawn from a standard normal distribution.

In this analysis, the dyadic clustering estimator and the exchangeable estimator are compared

in terms of the 95% confidence interval. This is repeated for $n = 10, 20, 40, 80$ with $R = 1$. For each sample size n, the covariates are then randomly generated 100 times. Subsequently, both error types are randomly simulated $1,000$ times.

The distribution of the exchangeable error can also be referred to as the bilinear mixed effects model (Hoff, 2005) and has previously been used in literature for justifying the use of the dyadic clustering estimator in simulation studies (Aronow et al., 2015).

Secondly, the generation of the non-exchangeable errors is done by means of adding a random effect with mean zero to only one quadrant of the covariance matrix, such that a reordering causes a different distribution.

Finally, the standard errors of the regression model are estimated using either the dyadic clustering or the exchangeable sandwich variance estimator. This will result in four confidence intervals for each combination of error settings and estimators for every n.

4.4 Empirical study

Next to the simulation study, the exchangeable estimator is also evaluated for real-world international trade data, as described in Section 3. The model to be estimated is given in equation 10 and possesses all the variables derived by Westveld & Hoff (2011). To assess the improvement originating from the exchangeable estimator, we include OLS as a benchmark in this analysis.

$$\begin{aligned}\log(\text{Trade}_{ijt}) = \beta_{1t} + \beta_{2t}\text{logGDP}_{jt} + \beta_{3t}\text{logGDP}_{it} + \beta_{4t}\text{logD}_{ijt} + \beta_{5t}\text{Pol}_{it} \\ + \beta_{6t}\text{Pol}_{jt} + \beta_{7t}\text{CC}_{ijt} + \beta_{8t}(\text{Pol}_{it} \times \text{Pol}_{jt}) + \epsilon_{ijt}\end{aligned} \tag{10}$$

The dependent variable, logTrade_{ijt}, represents the log volume of trade from country i to j in year $t = 1, ..., 20$. There model includes a constant β_{1t} and seven independent variables which are: logGDP_{it} and logGDP_{jt} which denotes the log GDP of countries i and j; logD_{ijt} is the distance between the countries in log; CC_{ijt} measures cooperation in conflict; and Pol_{it} and Pol_{jt} respectively indicate the polity of countries i and j, and are also used in a product.

4.4.1 Out-of-sample prediction

The two approaches, exchangeable and OLS, are compared by one-year-ahead forecasts for the years $t = 5, ..., 20$, i.e., the first point forecast is made at $t = 4$ for $t + 1 = 5$ by fitting the model on the first 4 years. While using an expanding window, the last forecast is made using 19 years of trade data for $t + 1 = 20$.

For OLS, the predictions are made with the assumption that there exists no auto-correlation. Hence, the variance-covariance matrix $\mathcal{E}_t$ is independent to $\mathcal{E}_{t+h}$ for any h. Additionally, the matrices $\mathcal{E}_t$ are identically distributed. This results in the following one-step-ahead estimate with OLS:

$$\mathbb{E}_{OLS}(y_{T+1}|\Im_T) = X_{T+1}\hat{\beta}_T, \tag{11}$$

where $\Im_T$ is the information set at time T, containing all available information at that time.

The exchangeable estimator rests on the assumption that the relations $\{y_t\}_{t=1}^{T}$ are joint normally distributed. Above that, some specifications are necessary. We denote the covariance structure between relations in the same relational context, in this case, years, by $\text{var}(y_t) = \Omega_1$ for $t = 1, ..., T$. When $r \neq s$, hence the year, is different, the covariance structure is denoted by $\text{cov}(y_t, y_{t+h}) = \Omega_2$ for all h. What follows is that the variance of the concatenated vector $z_{T-1} = (y_1, y_2, ..., y_{T-1})$, contains Ω_1 along the diagonal blocks and Ω_2 on the off-diagonal blocks. After taking the inverse of $\text{var}(z_{T-1})$, the diagonal blocks are denoted by Ψ_1 and the off-diagonals by Ψ_2. Based on this, the exchangeable estimate of y_{T+1} is constructed as follows:

$$\mathbb{E}_E(y_{T+1}|\Im_T) = X_{T+1}\hat{\beta}_T + \Omega_2(\Psi_1 + (T-1)\Psi_2)) \sum_{t=1}^{T}(y_t - X_t\hat{\beta}_t). \tag{12}$$

4.5 Clustering heterogeneity within the data

Furthermore, since Marrs et al. (2023) described that their proposed exchangeable estimator might not work as well when the data is heterogeneous, our paper suggests forming clusters within the data such that each clustered dataset holds the assumption of exchangeability. We propose two ideas to base the clusters on: one is empirical by $k-$means clustering and the second option is intuitive by clustering based on recognized connections.

After obtaining the clusters, the subgroups are each individually estimated by using the exchangeable estimator. The benchmark in this case is the exchangeable estimator when considering the complete unclustered dataset. To see if there has been an improvement with the additional clusters, the R^2 and the Mean Squared Prediction Error (MSPE) can be consulted.

4.5.1 $k-$means clustering

In order to partition the data into clusters, k-means clustering is performed as proposed by Hartigan & Wong (1979). The goal of this method is to minimise the within-cluster sum of squares (WCSS) such that optimal clusters are formed. In this paper, clustering is done based on the countries rather than based on the observations, for which clustering should be performed on the pairs. In the last case, clustering is less intuitive compared to clustering based on countries. Also, clustering on pairs is more difficult due to time-variant trade flows between the countries. The plus side to clustering based on countries is the easy interpretability. Additionally, we can cluster based on variables included in the dataset such as GDP and polity. However, we wish to incorporate a distance variable as well, which is not possible in country-based clustering.

The algorithm proceeds as follows: First, k points are randomly selected from the dataset as initial centroids. Second, all n observations are assigned to the nearest centroid such that the Euclidean distance from each observation to the centroid is minimal. This second step results in k clusters. Third, the cluster centroids are updated by taking the mean of all observations that lie within the cluster. Repeat steps two and three iteratively until convergence, which happens when the assignments of the cluster no longer change significantly, or when a maximum number of iterations is met. The final output of the $k-$means clustering algorithm are k cluster centroids with a set of observations assigned to each.

The downside to $k-$means clustering is the dependence on the initial random selection of

centroids. Therefore, the initial distribution can lead to different results. To overcome this, the algorithm is run multiple times to select the solution with the lowest WCSS.

The number of clusters can be based on either the Elbow method or interpretability. When investigating trade data, a prior analysis of the data can be done by identifying certain trade agreements, and based on this k can be chosen. The European Union is an example of a supranational organisation on which an interpretable selection of clusters can be formed.

4.5.2 F-test

The dataset contains 7 predictors, of which none of them take trade agreements into account. To see whether a European Union binary variable has a significant effect on the fit of the model, an F-test is conducted. The F-test compares two nested models, where one can be denoted as the smaller or the restricted model, and the other the larger or the unrestricted model. The null hypothesis suggests that both models perform equally well and it can be specified that the coefficients for the extra variables are equal to zero $\beta = 0$, or have no explanatory power. If the null is rejected, there is significant proof that the addition of the variable significantly improves the fit of the model.

To perform this test, first, both models are estimated with OLS. Based on this, the Sum of Squared Residuals can be computed for the restricted (SSR_r) and the unrestricted (SSR) model. Then, the following test statistic can be calculated:

$$F = \frac{(\text{SSR}_\text{r} - \text{SSR})/g}{\text{SSR}/(n-k)} \sim F(g, n-k). \tag{13}$$

where g is the number of restrictions, n the number of observations, and k the number of predictors in the full model. Under the null, the test statistic follows an F-distribution with degrees of freedom g and $n - k$.

4.5.3 Model comparison

Finally, the models' predictive abilities are measured in a two-fold analysis. The first encompasses the Mean Squared Prediction Error (MSPE) that calculates the difference between the real value, y_i, and the predicted value $\hat{y}_i$. The MSPE is calculated for the individual clusters to see if there is a gain from analysing a cluster on its own.

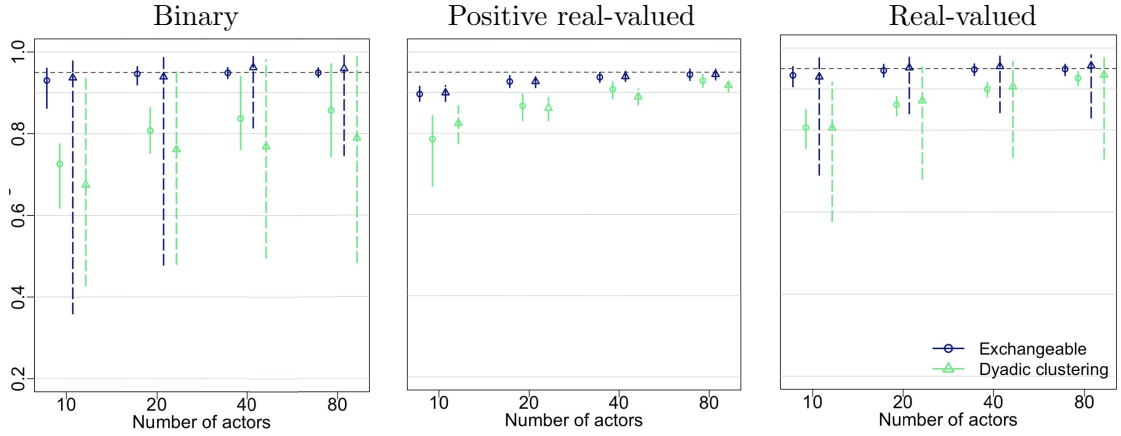
$$\text{MSPE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{14}$$

Then, after performing a series of F-tests to look for variables that significantly improve the model fit, the coefficient of determination, denoted by R^2, is consulted. R^2 describes the goodness of fit of a model and ranges from zero to one. These are compared for OLS, the exchangeable model and the clustered exchangeable model. The R^2 is computed in the following manner:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{15}$$

where $\bar{y}_i$ stands for the mean value of the dependent variable.

Figure 3: *Estimated coverages that the true coefficient falls within the 95% confidence interval.*



Note. the dark lines indicate the exchangeable estimator and the green lines indicate the dyadic clustering estimator when the errors are either generated from an exchangeable (circles) or non-exchangeable (triangles) distribution. The points represent the mean estimated coverage.

5 Results

The three analyses are now discussed. First, simulated data is used to compare the bias and variance of the exchangeable estimator against the dyadic clustering estimator. Secondly, real-world data enables forming and evaluation of predictions for the exchangeable model and ordinary least squares (OLS). Lastly, our contribution to the literature of a clustered exchangeable estimator is tested against one of Marrs et al. (2023).

5.1 Comparison of the exchangeable to the dyadic clustering estimator

In this analysis, the dyadic clustering estimator is compared to the exchangeable estimator. Since the latter is specified with the assumption of exchangeability, it is expected that it is more accurate when the data exhibits this property. On the other hand, the dyadic clustering estimator does not rest on this assumption. Therefore, we check for both cases when the data is exchangeable and when it is not. The dyadic clustering estimator contains more parameters which lead to a higher variance but a lower bias in the case of non-exchangeability.

In Figure 3, the simulation study is presented for binary, positive real-, and real-valued covariates, for $n = 10, 20, 40, 80$ and for a single relational context $R = 1$. The vertical lines stand for the estimated probability of the middle 95% of coverage intervals. When this interval is shorter, there is less variance in the prediction. The points, either circles or triangles, represent the estimated mean coverage. A point closer to the dashed 95% line indicates a lower bias. For each n, the first two lines, which are accompanied by a circle, show the simulation of data from an exchangeable dataset. The third and fourth lines, which are dashed and denoted by a triangle in each case, represent the non-exchangeable data.

In almost every instance, the coverage intervals for the exchangeable data are smaller than those for the dashed lines of the non-exchangeable data, which is easily clarified by the more restrictive nature of exchangeable data. Also, across all three figures, the blue points, thus the mean coverage of the exchangeable estimator, lie closer to the 95% level than the green points. This means that the exchangeable estimator is less noisy than the dyadic clustering estimator

in all cases.

Especially in the left panel, where the binary data is given, the exchangeable estimator is much closer to the nominal level of 0.95. This is due to the noisiness of binary data. In comparison to the paper of Marrs et al. (2023), almost all coverage intervals are the same except for the exchangeable estimator for non-exchangeable data. In our case, the length of those intervals is longer. This can be explained through the fact that we work with a simulation study, which could cause different means across simulations. Additionally, the difference becomes more pronounced for binary data.

The middle panel represents the positive real-valued data. For both types of simulations, the exchangeable estimator performs better while increasing n. This is exactly in accordance with Marrs et al. (2023).

The last panel contains real-valued data and is similar to the seminal paper of our research except that the lengths of the estimated confidence intervals of the non-exchangeable data are a bit shorter in our plots. Again, we can account for this difference by simulation.

5.2 Out-of-sample prediction study

Real-world data is used to evaluate the fit of the exchangeable estimator compared to OLS, which uses dyadic clustering standard errors, as a benchmark. We have international trade data of $n = 58$ countries over a period of $R = T = 20$ years, which has been derived by Westveld & Hoff (2011). By means of an out-of-sample prediction study the two models are compared with one-year-ahead forecasting. The forecasts will be made for $t + 1 = 5, ..., 20$. The goodness of fit can be analysed with the coefficient of determination R^2, which measures the proportion of the variance that can be explained by the dependent variables.

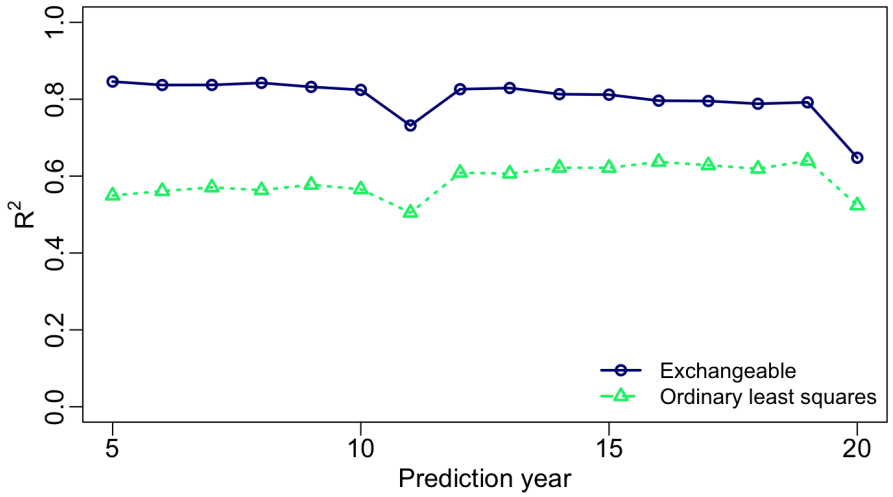
In Figure 4, the R^2's are plotted against the prediction years. The blue line represents the R^2 over the out-of-sample window for the exchangeable estimator and the green line shows the R^2 for the benchmark, OLS. It is clear that the exchangeable approach yields a better fit for the model. The mean R^2, in this case, is equal to 0.80, while OLS has a mean of 0.59. In the beginning, this difference in R^2 is larger (around 0.30), but for the higher t, the difference declines. Especially in 2000 ($t = 20$) the difference is only 0.12, meaning that there is more equal performance. The very apparent drop in fit for both models from 1999-2000 could be explained by the establishment of the Euro by the European Monetary Union on January 1st, 1999. Similarly, in 1991 ($t = 11$) there is a visible drop as well. The cause of this can be clarified by the end of the Cold War. In 1991 the Soviet Union dissolved and caused the end of a period of reduced trade due to barriers. After this, global economies were enhanced again but first caused worse predictability.

We notice an exact resemblance to the results of Marrs et al. (2023).

5.3 Clustered exchangeability

We want to improve upon the exchangeable estimator proposed by Marrs et al. (2023) by separately estimating the clusters that are present within the data. To this end, we commence with extracting clusters from the data. $k-$means clustering optimizes similarity within a cluster by minimizing the distance between all the data points and the cluster's center. At the same

Figure 4: R^2 *of the exchangeable estimator and OLS*



Note. the one-year-ahead predictions for $t+1 = 5, ..., 20$, for the exchangeable estimator (blue circles) and OLS (green triangles).

time, dissimilarity is maximized between clusters. Clustering is based on the countries, and not on the pairs, hence only relevant variables on the exporting country are included. These are the standardised variables logGDP_i and Pol_i per country $i = 1, .., 58$. Additionally, the variables are averaged over the time periods to obtain constant clusters.

Initially, we perform this analysis for $k = 2, 3, 4, 5$ as seen in Figure 6 in the Appendix. It follows that at first, the most important divider is GDP. This Figure is also in line with the heavy-tailed distribution of the polity of the country, as was visible in Figure 2.

Interestingly, all European Union member states (except for Cyprus, which was not yet a member state during the prediction window) lie in the upper right quadrant. These countries also remain in the same cluster (cluster 2) until $k = 3$. As a result, we examine the explanatory power of a binary variable indicating whether a country is a member state of the EU. Since several countries joined the EU during 1980-1995, we limit our analysis to 1995-2000.

Two linear regressions are performed: (1) OLS on the original model, (2) OLS on the original model plus an extra binary variable, which is equal to 1 if both the exporting country i and importing country j are member to the EU. The results from these regressions are presented in Table 2. All variables are highly significant (p-value < 0.000) meaning that each variable contains valuable information to the model. It follows that countries with higher GDPs, both exporting or importing, trade more. Moreover, the distance between countries negatively influences the amount traded between them. Particularly, the larger model has a significant but negative coefficient for the extra EU variable, which can partially be explained by the size of the economies.

The R^2 for the small and the large model are respectively 0.612 and 0.614, hence there is only a two-thousandths increase. The partial F-test is consulted to check for a significant increase in the coefficient of determination. The test statistic is equal to 116.919 with a p-value of < 0.000, suggesting that the model with the additional EU variable provides a better fit. This implies that the variable should be included in the model and therefore we can rectify the decision to

Table 2: *The regression results including a binary EU variable*

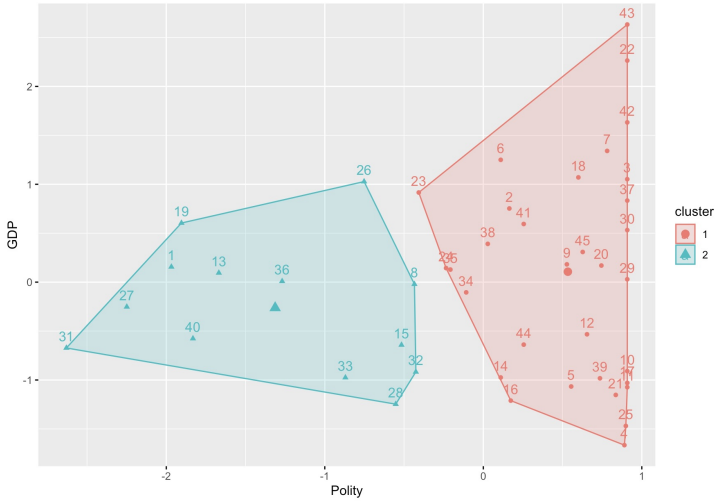
	Original Model			**Model with EU**		
	Estimate		P-value	Estimate		P-value
Intercept	-51.120	(0.444)	0.000*	-51.690	(0.446)	0.000*
logGDP_i	1.464	(0.012)	0.000*	1.477	(0.012)	0.000*
logGDP_j	1.194	(0.012)	0.000*	1.207	(0.012)	0.000*
logD_{ij}	-1.410	(0.026)	0.000*	-1.516	(0.028)	0.000*
Pol_i	0.088	(0.005)	0.000*	0.093	(0.005)	0.000*
Pol_j	0.097	(0.005)	0.000*	0.101	(0.005)	0.000*
CC_{ij}	-0.647	(0.059)	0.000*	-0.523	(0.060)	0.000*
EU_{ij}				-1.290	(0.119)	0.000*
R^2	0.612			0.614		

Note. On the left, the coefficients for the variables of the original model with the standard deviation in brackets and the corresponding p-value are given. The right model represents the model with the extra binary EU variable. At the bottom, the R^2's are given.

form an EU cluster. By first splitting the dataset into EU and non-EU, the distance variable can partially be taken into account, which could otherwise not happen based on the clustering of individuals.

The EU cluster provides a starting point for our further clustering analysis. Based on the 45 countries that remain, $k-$means clustering is again performed. The following two clusters are formed and given in Figure 5. The first cluster, now denoted by cluster A, can be characterised by low levels of polity and relatively low levels of GDP. The second cluster, now called cluster B, has a wider variety of GDP levels but high polity. In Table 6 in the Appendix, the complete list of countries and which cluster they are in is given.

Figure 5: *Clusters of the non-EU countries*



The summary statistics for the clusters are given in Table 3. The three columns denoted by cluster indicate within cluster trade, meaning that both the importing and exporting countries need to be in the same cluster. Noticeable is the low levels of trade in cluster A. By looking at Table 6, cluster A contains some Mid-American countries or countries that are far away from each other and unrelated. This is confirmed by investigating the data since the traded amount between several pairs of countries is equal to zero during multiple time periods. In comparison to

Table 3: *Summary statistics for the full sample, clusters, and cross-clusters*

	Original	**Cluster**			**Cross clusters**		
		EU	A	B	EU_A	EU_B	A_B
Trade	15.721	21.423	11.576	16.295	17.366	18.322	13.852
Polity	0	0.717	-1.660	0.308	-0.164	0.532	-0.275
GDP	0	0.690	-0.485	-0.118	0.197	0.418	-0.177

the full sample, denoted by the original model, trade is higher for clusters EU and B. Especially the EU has a large volume of trade flow. This is in line with its GDP, which is positively correlated with traded volume. Note that the mean variables for polity and GDP for the full sample are equal to zero due to standardisation. Overall, cluster B has averages lying closest to the mean value of the full sample.

Additionally, we are interested in the cross-cluster data for the estimation of the model. Similar to the covariance estimation of relational arrays, we assume the parameter ϕ_{AB} of the directed array y_{ij} where $i \in A, j \in B$ to be the same as ϕ_{BA}. Hence, cross-cluster AB contains both directions. Again, in cluster AB, several trade flows are equal to zero, leading to a low average of trade flow in Table 3.

In accordance with the (cross-)clusters, we form six new datasets, after which each model is separately estimated. First, the coefficients are estimated, analogously to the exchangeable method as described in the previous section. Then, the one-step-ahead predictions can be made. To sufficiently train the model years 1995-1998 are used as an in-sample. The years 1999-2000 are out-of-sample and are predicted one-step-ahead. These two years are also interesting due to the drop in R^2 that was visible in Figure 4.

By separating the data into clusters, different problems are encountered. Cluster A only contains zeroes for the conflict variable, hence this column causes a singular matrix. Therefore, to preserve fairness in comparison, the variable conflict is omitted. Additionally, the polity causes linearly dependent columns for the EU dataset. This is due to the fact that almost all polity values are equal to 20 for both the importing and exporting countries. To erase problems caused by this, dependent columns can be combined into one: a product of the two values.

Finally, to see if there has been an improvement from incorporating clusters, we compare the predicted value to the observed value by means of the Mean Squared Prediction Error (MSPE). The results from the estimation are given in Table 4. The first column shows the MSPE of the original model, containing no clusters. The full dataset has 6 years of observations of $58 * 57$ pairs. The MSPE is clearly higher for the year 2000 than in 1999, which was visible in Figure 4. Moreover, the average MSPE of these two years is equal to 7.095.

Interestingly, the predictive performance of the EU model is increased. The clustered exchangeable estimator is able to make accurate forecasts and shows it is profitable to separately estimate this cluster. On the contrary, cluster A behaves extremely poorly and explains a major part of the worse behaviour overall. Cluster B creates a larger bias on average for the year 1999 than the full sample, but outperforms the full sample in terms of MSPE in 2000.

In comparison to the average of the weighted MSPEs, the clustered exchangeable estimator creates a lower bias than the unclustered one. However, this gives a distorted image of predictability since the cross-clusters are not considered. It is plausible reasoning to assume that

Table 4: *MSPEs for the original model and the clusters*

	Original	**Cluster**		
	full	EU	A	B
Observations	19836	936	792	5952
MSPE_{1999}	4.680	0.432	35.002	8.392
MSPE_{2000}	9.509	0.280	26.499	5.554
Mean MSPE	7.095	0.356	30.723	6.973
Weighted average	7.095		7.071	

forecasting trade flows from a country from one cluster to a country in another is more arduous since there is less similarity. Yet, since the weighted average is mainly influenced by the high MSPE of cluster A, it is unsure how much harder this actually is. Cluster A is characterised by unrelated countries, therefore, this might just as well be the hardest to predict out of the cross-clusters as well.

The final approach to testing whether incorporating clusters adds value to the model is by comparing the R^2. In Table 5, the R^2's are given for the exchangeable estimator as we determined in 5.2 for the last three years of the sample. In Figure 4, there was a clear drop in model fit in the last year, which could be explained by emerging differences between countries. The euro was implemented in 1999 and started having its global effect. Secondly, the burst of the dotcom bubble polarised more- and less-advanced countries. Additionally, there were numerous financial crises in the late 1990s which intensified differences between suffering and thriving states. In consequence, incorporating clusters is able to explain fluctuations.

There is a noticeable difference when comparing the values of the original model to the exchangeable model that includes binary variables for the clusters. The increase to the R^2 when the EU variable is added amounts to 0.18. Therefore, the F-test yields a significant difference. Secondly, we add the cluster A variable since this cluster deviates the most from the mean and therefore could have explanatory power to outlying observations. There is a slight increase in comparison to the model with only EU, yet, it is a significant increase with p-value < 0.000. Finally, we form a model with all three clusters. When investigating the decimals it turns out that including B yields a lower R^2. Thus, the best option for predicting is to include the EU and A binary variables. In general, there is a significant gain from adding clusters to the exchangeable model when investigating trade flows.

Table 5: R^2 *for the exchangeable and the clustered exchangeable estimator*

	Exchangeable	**Clustered exchangeable**		
		EU	EU and A	EU, A and B
R^2_{1998}	0.788	0.848	0.850	0.850
R^2_{1999}	0.791	0.855	0.856	0.856
R^2_{2000}	0.648	0.868	0.870	0.870

Note. the last three columns respectively show the R^2 of the model with the EU binary variable, the EU and A variable, and the model with all three binary variables for the years 1998-2000.

6 Conclusion

In this research, the main objective is to develop a regression framework for predicting exchangeable relational arrays while leveraging the exchangeability property to obtain a symmetric and parsimonious covariance matrix. This yields a covariance matrix that contains at most ten parameters to be estimated. To extend the state-of-the-art literature, we apply this framework to the same dataset but acknowledge the heterogeneity within the data by extracting clusters, both empirically and intuitively.

First of all, the exchangeable estimator was compared to the dyadic clustering estimator in an analysis of confidence intervals with simulated data. The results were unanimous and in favour of the exchangeable estimator even when the data was non-exchangeable. When analysing the fit of the exchangeable and ordinary least squares model for international trade data, there was again clear evidence that the exchangeable approach was beneficial. When including the clusters, dubious results followed. For the European Union cluster, it followed that it is advantageous to estimate and predict the model only using the pairs which were in the EU.

The most important contribution to the research question of how to best model exchangeable relational data when there is heterogeneity in the data is by incorporating binary variables for the clusters in the regression. This leads to the conclusion that investigation and inclusion of clusters within the data are lucrative to predicting trade flows one-year-ahead.

However, there are some limitations. The decision to cluster based on countries also has a downfall since it does not incorporate trading treaties or distances, even though the EU cluster was included. Furthermore, we initiated constant clusters for simplicity, yet, time-varying clusters over larger samples are wiser since countries are developing. Above that, for ease of deductibility, we only used a small number of clusters. This leads to the fact that cluster A contains a wide variety of unrelated nations. Lastly, it would be wise to cluster based on more information, since polity and GDP are only two influential variables. A possible suggestion for further research would be to also cluster based on the trade between countries, since in our case sometimes there were no trade flows between countries i and j even though they were in the same cluster.

Based on these limitations, we suggest further enhancing a clustering algorithm that optimises the data available and makes sure the countries in each cluster are in fact affiliated.

Even though our research focuses on international trade data, the clustered exchangeable estimator can be applied to a wider variety of datasets. The exchangeable estimator provides a parsimonious way of estimating relational arrays that are better explained by an additional set of clusters. To finalise, our research provides the building blocks for incorporating clusters in the setting of relational arrays that leverage exchangeability. Consequently, predictability is enhanced and more profound structures within the data can be understood.

References

Aldous, D. J. (1981). Representations for partially exchangeable arrays of random variables. *Journal of Multivariate Analysis*, *11*(4), 581–598.

Aronow, P. M., Samii, C. & Assenova, V. A. (2015). Cluster–robust variance estimation for dyadic data. *Political Analysis*, *23*(4), 564–577.

Attanasio, O., Barr, A., Cardenas, J. C., Genicot, G. & Meghir, C. (2012). Risk pooling, risk preferences, and social networks. *American Economic Journal: Applied Economics*, *4*(2), 134–167.

Beck, N., Gleditsch, K. S. & Beardsley, K. (2006). Space is more than geography: Using spatial econometrics in the study of political economy. *International studies quarterly*, *50*(1), 27–44.

Cai, D., Ackerman, N. & Freer, C. (2016). Priors on exchangeable directed graphs.

Carlson, J., Incerti, T. & Aronow, P. (2021). Dyadic clustering in international relations. *arXiv preprint arXiv:2109.03774*.

Chen, J., Song, Q., Zhao, C. & Li, Z. (2020). Graph database and relational database performance comparison on a transportation network. In *Advances in computing and data sciences: 4th international conference, icacds 2020, valletta, malta, april 24–25, 2020, revised selected papers 4* (pp. 407–418).

Conley, T. G. (1999). Gmm estimation with cross sectional dependence. *Journal of econometrics*, *92*(1), 1–45.

Crane, H. & Dempsey, W. (2019). Relational exchangeability. *Journal of Applied Probability*, *56*(1), 192–208.

Dicken, P., Kelly, P. F., Olds, K. & Wai-Chung Yeung, H. (2001). Chains and networks, territories and scales: towards a relational framework for analysing the global economy. *Global networks*, *1*(2), 89–112.

Fafchamps, M. & Gubert, F. (2007). The formation of risk sharing networks. *Journal of development Economics*, *83*(2), 326–350.

Fan, X., Li, Y., Chen, L., Li, B. & Sisson, S. A. (2020). Smoothing graphons for modelling exchangeable relational data. *arXiv preprint arXiv:2002.11159*.

Fosdick, B. K. (2013). *Modeling heterogeneity within and between matrices and arrays* (Unpublished doctoral dissertation).

Fosdick, B. K. & Hoff, P. D. (2014). Separable factor analysis with applications to mortality data. *The annals of applied statistics*, *8*(1), 120.

Hartigan, J. A. & Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the royal statistical society. series c (applied statistics)*, *28*(1), 100–108.

Hoff, P. D. (2003). *Random effects models for network data.* na.

Hoff, P. D. (2005). Bilinear mixed-effects models for dyadic data. *Journal of the american Statistical association*, *100*(469), 286–295.

Hoff, P. D. (2009). Multiplicative latent factor models for description and prediction of social networks. *Computational and mathematical organization theory*, *15*(4), 261–272.

Hoff, P. D. & Ward, M. D. (2004). Modeling dependencies in international relations networks. *Political Analysis*, *12*(2), 160–175.

Huber, P. J. (1967). Under nonstandard conditions. In *Proceedings of the fifth berkeley symposium on mathematical statistics and probability: Weather modification; university of california press: Berkeley, ca, usa* (p. 221).

Lloyd, J. R., Orbanz, P., Ghahramani, Z. & Roy, D. M. (2013). Exchangeable databases and their functional representation. In *Nips workshop on frontiers of network analysis: Methods, models, and application.*

Long, B., Zhang, Z. & Philip, S. Y. (2010). *Relational data clustering: models, algorithms, and applications.* CRC Press.

Long, B., Zhang, Z. M. & Yu, P. S. (2007). A probabilistic framework for relational clustering. In *Proceedings of the 13th acm sigkdd international conference on knowledge discovery and data mining* (pp. 470–479).

Marrs, F. W., Fosdick, B. K. & McCormick, T. H. (2023). Regression of exchangeable relational arrays. *Biometrika*, *110*(1), 265–272.

Walsh, B., Mohamed, S. K. & Nováček, V. (2020). Biokg: A knowledge graph for relational learning on biological data. In *Proceedings of the 29th acm international conference on information & knowledge management* (pp. 3173–3180).

Warner, R. M., Kenny, D. A. & Stoto, M. (1979). A new round robin analysis of variance for social interaction data. *Journal of Personality and Social Psychology*, *37*(10), 1742.

Westveld, A. H. & Hoff, P. D. (2011). A mixed effects model for longitudinal relational and network data, with applications to international trade and conflict.

Zahay, D., Peltier, J., Schultz, D. E. & Griffin, A. (2004). The role of transactional versus relational data in imc programs: Bringing customer data together. *Journal of advertising research*, *44*(1), 3–18.

A Appendix

The appendix comprises additional figures and tables as well as a more in-depth explanation of the alterations made to the data and code.

Table 6: *The list of countries and the clusters in which they belong*

	Countries	Cluster				Countries	Cluster		
		EU	A	B			EU	A	B
1	Algeria		x		*30*	Jamaica			x
2	Argentina			x	*31*	Japan			x
3	Australia			x	*32*	Republic of Korea			x
4	Austria	x			*33*	Malaysia			x
5	Barbados			x	*34*	Mauritius			x
6	Belgium	x			*35*	Mexico		x	
7	Bolivia			x	*36*	Morocco		x	
8	Brazil			x	*37*	Nepal		x	
9	Canada			x	*38*	The Netherlands	x		
10	Chile			x	*39*	New Zealand			x
11	Colombia			x	*40*	Norway			x
12	Costa Rica			x	*41*	Oman		x	
13	Cyprus			x	*42*	Panama		x	
14	Denmark	x			*43*	Paraguay		x	
15	Ecuador			x	*44*	Peru			x
16	Arabic Republic & Egypt		x		*45*	Philippines			x
17	El Salvador			x	*46*	Portugal	x		
18	Finland	x			*47*	Singapore		x	
19	France	x			*48*	Spain	x		
20	Germany	x			*49*	Sweden	x		
21	Greece	x			*50*	Switzerland			x
22	Guatemala			x	*51*	Thailand			x
23	Honduras		x		*52*	Trinidad and Tobago			x
24	Iceland			x	*53*	Tunisia		x	
25	India			x	*54*	Turkey			x
26	Indonesia		x		*55*	United Kingdom			x
27	Ireland	x			*56*	United States			x
28	Israel			x	*57*	Uruguay			x
29	Italy	x			*58*	Venezuela			x
	Total						13	12	33

A.1 Data

In order to use the data derived by Westveld & Hoff (2011), some alterations were necessary. Since the delimiter in this case was an empty space, countries such as El Salvador are to be changed to ElSalvador. Furthermore, the column names contain points as separators, which should be altered to spaces. Above that, the first column is unnecessary and quotation marks can be deleted.

Then, as the clustering is based on the polity and GDP, these should be standardized to have equal influence. Then, the mean is also calculated over 20 years of time. For the clustering

analysis, a careful approach should be chosen to deal with the right indexes for the countries. Since we treat the EU separately, these indexes are saved in a vector. Then, the clustering analysis is performed for which it is important to use the same ordering, excluding the EU countries. Noteworthy is that the countries are not ordered alphabetically.

Lastly, for the analysis including the EU, the years 1995-2000 are considered. Therefore, indexes 15-20 should be changed to 1-6 in order to maintain the code's usefulness. This is also necessary for the clusters and cross-clusters.

A.2 Code

In this section, we mention some important explanations for the code.

Marrs et al. (2023) has made a significant contribution to the code used in this paper. Therefore, important alterations are mentioned. In the simulation study, exchangeable and non-exchangeable errors are simulated. The authors of the code use the

Furthermore, for each implemented dataset, with either a single cluster or the full model with binary variables, new directories, and results should be formed accordingly such that the estimates β's are put in the right place and the estimation is based on the correct set. Additionally, with every other set, the amount of nodes changes. When the fits are formed for the model with binary variables, this means an extra variable is included such that the size of X increases. Therefore, β also enlarges and the number of columns should be modified. It is important that when switching to another dataset, the environment is cleaned.

When fitting the clustered models individually, we deal with singularity issues. Since the column of conflict only contains zeroes for cluster A, this column should be omitted such that the columns are not linearly dependent. A similar issue happens for the cross-cluster EU_B since the polity has the same value for $t = 3$. The singularity occurs due to the strongly correlated variables especially in the EU cluster due to more similarity.

Figure 6: *k−means clustering for various k for the whole dataset*

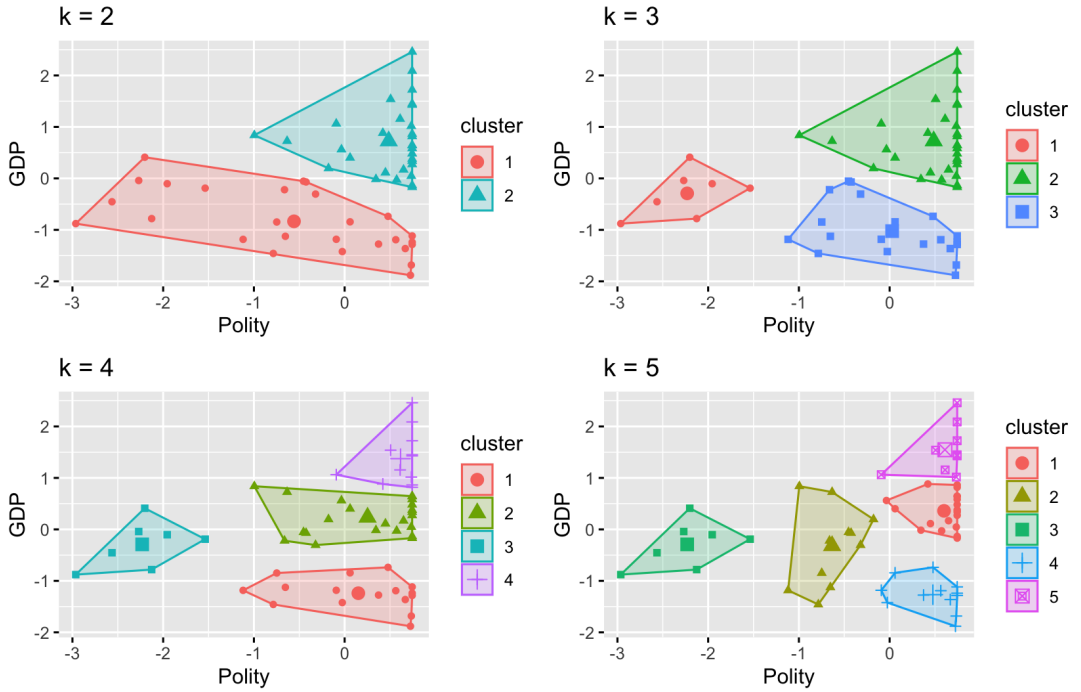


Figure 7: *The countries and where they are positioned in the GDP vs. polity spectrum*