Journal of Statistical and Econometric Methods, vol.7, no.3, 2018, 63-87
ISSN: 2241-0384 (print), 2241-0376 (online)
Scienpress Ltd, 2018

A method for clustering panel data based on parameter homogeneity

Juan Romero-Padilla[1]

Abstract

Panel data models assume that parameters are common to each subject, that assumption is not satisfied in many cases. The slope heterogeneity problem may be solved by obtaining groups where the slope parameters are heterogeneous across groups but homogeneous within groups, followed by panel data theory within each group. In this paper, an algorithm to determine clusters of subjects is discussed; the clustering is achieved by checking whether confidence intervals from different subjects overlap or not. The number of groups is determined based on the data variability. The clusters are useful by themselves to analyze the similar behavior of subjects. Monte Carlo simulations were performed to examine the properties of the methodology considered. Finally, clusters of countries with similar GDP per capita trend were obtained.

Mathematics Subject Classification: 62F03, 62F25, 62H30, 91G70

Keywords: Panel data models, Clustering of subjects, Parameter homogeneity test, Confidence intervals

[1] Center for Research and Teaching in Economics (CIDE), LNPP-Conacyt. Mexico City, Mexico.

Article Info: Received: December 6, 2018. Revised : January 22, 2018
Published online : July 15, 2018

1 Introduction

Traditional panel data models assume parameter homogeneity; such assumption is rejected in some empirical studies ([1], [2], [3]). One approach to deal with slope heterogeneity in panel data models is, to obtain groups where the parameters are heterogeneous across groups but homogeneous within groups; under this approach, Su et al. [4] proposed a variant of Lasso to estimate heterogeneous linear panel data; Su et al. [5] proposed a heterogeneous time-varying panel data model with latent group structures to capture individual heterogeneity; Lin and Ng [6], proposed two methods to obtain groups, the first one was a modification of the K-means algorithm and the second one was built based on time series estimations for individual slope coefficients. In all the papers above mentioned they performed a priory estimation of the number of clusters. Hoogstrate et al. [7] analyzed pooling models when the parameters are different but exhibit some similarity, they used F-tests to decide when to pool. Frühwirth-Schnatter et al. [8] used Bayesian framework to cluster multiple time series, the estimations of their models were obtained using Markov Chain and Monte Carlo methods, the combination of those two techniques may not be familiar to some applied econometricians.

In this paper, we propose a methodology to determine clusters of subjects based on confidence intervals of parameters, the methodology follows an algorithm to assign subjects to specific clusters; for each subject, the parameter and confidence interval are estimated using a general linear model; if the confidence intervals of two subjects overlap, then we assume that the two subjects may belong to the same cluster.

This paper is organized as follows, some traditional panel data models are reviewed in section 2, the clustering panel data methodology is developed in section 3, classifications performance and estimations of the proposed clustering method are examined in section 4, section 5 provides an empirical study on the heterogeneous trending behavior of GDP per capita across countries and section 6 concludes.

2 Econometric Framework

A cross-sectional model is given by ([9], [10])

$$y_{it} = \beta_0 + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \qquad i = 1,2,\dots,N;\ t = 1,2,\dots,T \tag{1}$$

where β_0 is a scalar, $\boldsymbol{\beta}$ is a (k x 1) vector of parameters associated with k explanatory variables, $\boldsymbol{x}'_{it}$ is a vector of explanatory variables and ε_{it} is the error term. Subscript i denotes cross-section dimension whereas t denotes time-series dimension. The representation that uses information of repeated measurements on a subject is the fixed effects model

$$y_{it} = \beta_{0i} + \boldsymbol{x}'_{it}\boldsymbol{\beta} + \varepsilon_{it}, \tag{2}$$

where β_{0i} are allowed to vary by subject, β_j $(j = 1,2,\dots,k)$ are common to each subject and are called global, the parameters β_{0i} are non-estimable in cross-sectional regression models without repeated observations.

If individuals are randomly selected from a population, is more reasonable to represent β_{0i} as a random variable instead of fixed. If in equation (2), the term β_{0i} is assumed to be a random variable, then equation (2) represents the random effect model. The random effect model is a special case of the mixed linear model because includes random effects (β_{0i}) and fixed effects ($\boldsymbol{\beta}$).

There are several hypothesis tests to verify model assumptions and to decide between random model and fixed model, we describe briefly the tests used in this paper, for details see Baltagi [9] chapter 4; Greene [11] chapter 9; Fress [10] chapters 3, 4 and Hausman [12].

Hypothesis test to check if the same coefficients apply to each subject

$$H_0: \beta_{ik} = \beta_k^0 \ \textit{for all } i \text{ vs } H_1: \beta_{ik} \neq \beta_k^0 \ \textit{for some } \ i \tag{3}$$

Hypothesis test to check if Ordinary Least Squares (OLS) is better than fixed effects, i.e. no panel effect

$$H_0: \beta_{0i} = \beta_0 \ \textit{for all } \ i \ \text{ vs } \ H_1: \beta_{0i} \neq \beta_0 \ \textit{for some } \ i \tag{4}$$

Hypothesis test to check if OLS is better than random effects, i.e. no panel effect

$$H_0\colon \sigma^2_{\beta_{0i}} = 0 \text{ vs } H_1\colon \sigma^2_{\beta_{0i}} \neq 0 \tag{5}$$

Hypothesis test to decide between fixed effects model and random effects model

H_0*: Preferred model is random effects vs* H_1*: Preferred model is fixed effects* (6)

A standard F test is used to test slope homogeneity (3) and fixed effects (4), the F test is based on the comparison of a model obtained for the full sample and a model based on the estimation of an equation for each subject. To decide between a random effects regression and a simple OLS regression (5), the Breusch-Pagan Lagrange multiplier (LM) test is used, the null hypothesis in the LM test is that variances across entities are zero. To test the hypothesis (6) we use Hausman test, Hausman's result is that the covariance of an efficient estimator with its difference from an inefficient estimator is zero.

Before applying panel data theory is necessary to check if the assumptions are met, tests (4) and (5) are used to decide if there is panel effect. Furthermore, it is necessary to check if the same coefficients apply to each subject, otherwise, we need to explore alternative methods or consider clusters where the hypothesis test (3) is not rejected. All the tests mentioned here are available in R package plm, version 1.6-5, developed by Croissant and Millo [13].

3 Clustering Panel Data Methodology

We use similar notation to the one used by Lin and Ng [6], to describe a clustering model, the panel data model is rewritten as follows

$$y_{it} = \beta_{0i} + \boldsymbol{x'_{it}\beta_{(i)}} + \varepsilon_{it}; \qquad i = 1,2,\dots,N;\ t = 1,2,\dots,T \tag{7}$$

where $\boldsymbol{x'_{it}} = (x_{it1}, \dots, x_{itk})$ is a (k × 1) vector of explanatory variables, $\boldsymbol{\beta}_{(i)} = (\beta_{1(i)}, \dots, \beta_{k(i)})'$ is a (1 × k) vector of slope coefficients for subject (a_i), β_{0i} is the unobserved heterogeneity and ε_{it} is the error term.

If we introduce groups or clusters, model (7) becomes

$$y_{it} = \beta_{0i} + \boldsymbol{x'_{it}\beta_g} + \varepsilon_{it}; i \in G_g; \qquad g = 1,\dots,G \tag{8}$$

where G_g is an indicator variable for true group membership, G is the number of clusters, $\beta_g = (\beta_{1g}, \dots, \beta_{kg})'$ is a (1×k) vector of group-specific slope coefficients such that, for a given *k,* $\beta_{(i)}$ equals or is well approximated by β_g for all i's in G_g; N_g denote the number of subjects in cluster G_g with $\sum_{g=1}^{G} N_g = N$.

The clustering methodology is based on three assumptions.

Assumption 3.1 *Two subjects may belong to the same cluster if their parameters are statistically equal.*

To decide if two parameters are statistically equal we use the following hypothesis test

$$H_0: \beta_{k(i)} = \beta_{k(j)}\ vs\ H_1: \beta_{k(i)} \neq \beta_{k(j)} \tag{9}$$

If the null hypothesis is rejected, then the two subjects do not belong to the same cluster otherwise they may belong to the same cluster. To test (9) a confidence interval is used, if the value of the parameter specified by the null hypothesis is contained in the $(1 - \alpha)100\%$ confidence interval then the null hypothesis cannot be rejected at the α level, otherwise, it can be rejected at the α level.

Assumption 3.2 *A cluster is well defined if there is parameter homogeneity within the cluster.*

To decide if there is parameter homogeneity within each cluster we use the following test.

$$H_0: \beta_{k(i)} = \beta_{kg} \text{ for all } i \in G_g \quad \text{vs} \quad H_1: \beta_{k(i)} \neq \beta_{kg} \text{ for some } i \in G_g \tag{10}$$

Test (10) is equivalent to test (3) and both hypotheses are tested in the same way. Furthermore, Su and Chen [2] proposed a test of homogeneity in panel data models with interactive fixed effects.

Assumption 3.3 *A group of clusters obtained from N subjects is well defined if, for each cluster, assumption 2 is fulfilled.*

Next, the methodology to find the cluster's membership (G_g) is described. Without loss of generality, one explanatory variable is analyzed. First, the parameters ($\beta_{(i)}$) and confidence intervals ($L_{(i)}$,$U_{(i)}$) are estimated by OLS model. For a fixed subject $a_{(i)}$, there is a group $A_{(i)}$ that contains $n_{(i)}$ subjects for which test (9) is not rejected ($n_{(i)} \leq N$). The set $A_{(i)}$ is a candidate to be a cluster. However, if two elements of $A_{(i)}$ are chosen, say $a_{(m)}$ and $a_{(n)}$ with $i \neq m \neq n$, the hypothesis $H_0 \colon \beta_{(m)} = \beta_{(n)}$ may be rejected. If for subject $a_{(m)}$, test (9) is rejected in $n_{(i)} - 1$ comparisons then subject $a_{(m)}$ must be removed from set $A_{(i)}$, in general, subjects for which test (9) is rejected in more than $(n_{(i)} * q)$ comparisons must be removed ($0 < q < 1$). For example, if $q = 0.5$ then we remove subjects for whom test (9) is rejected in half or more comparisons. If q is close to one then $A_{(i)}$ will have elements with a full confidence interval overlap. The cluster assignation depends on the selection of the initial subject, because of that, the subject associated with the maximum $n_{(i)}$ is selected as initial subject. Once $A_{(i)}$ is defined their elements are classified and the process continues with the remained subjects until every subject has been assigned to a cluster. Follow, a test for parameter homogeneity within each cluster is carried out, if the hypothesis (10) is not rejected, the obtained clusters are correct, otherwise, it is necessary to increase q and repeat the process. In our results, it was found that when q is close to one, the number of clusters increases, therefore we recommend to start with $q = 0.3$ and increase it by 0.1 if it is necessary.

The above process is carried out for each explanatory variable, follow the different clusters are sorted as shown in Figure 1, the most important variable must be Var1, the second most important variable must be Var2 and so on. The maximum number of clusters to be obtained is $\mathrm{n}_{\max} = \left(n_{(1)}\right)\left(n_{(2)}\right) \cdots \left(n_{(N)}\right)$. In the proposed methodology can be inferred that if the number of explanatory variables increases then the number of clusters increases.

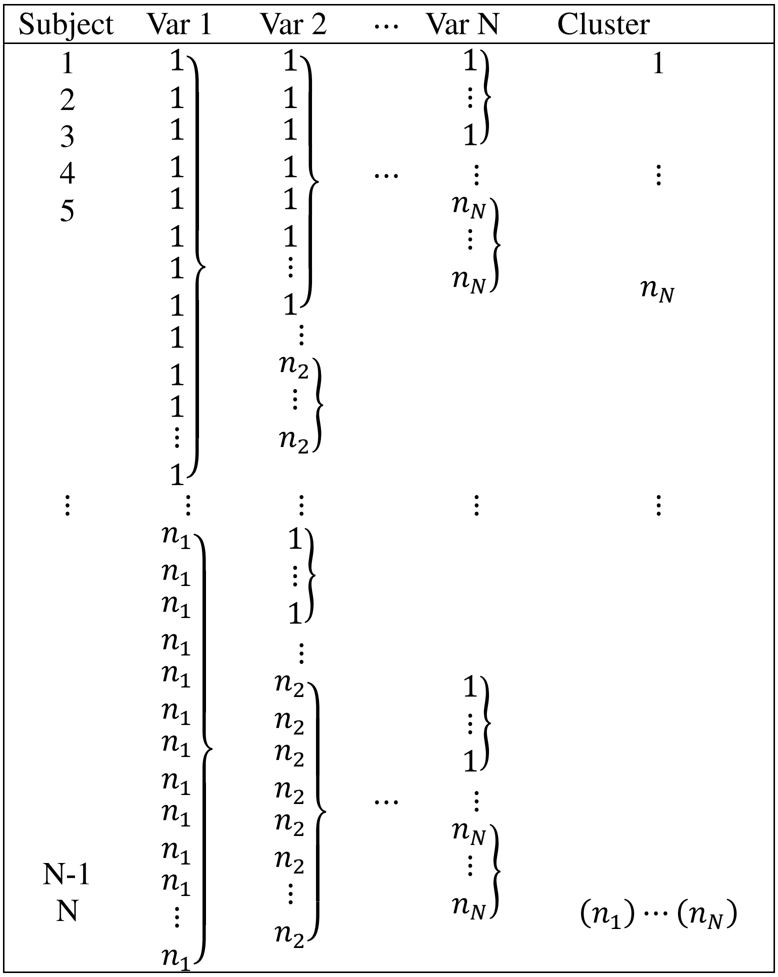

Figure 1: Order of variables to obtain the final clusters

3.1 Clustering Algorithm

Clusters are determined by the following clustering algorithm.

1. Check assumption 3, if hypothesis (10) is rejected for some cluster, then follow to step 2, otherwise stop the process. (At the beginning $G = 1$)
2. For each explanatory variable made clusters as follow
    a. For each subject $a_{(i)}$, using OLS obtain estimations of parameters and $(1-\alpha)100\%$ confidence intervals, denote $\beta_{(i)}$ and (L_i, U_i) as the parameter and confidence interval associated to $a_{(i)}$.

b. Define a matrix $\boldsymbol{A}$ (N x N) with elements defined as, $a_{ij} = 1$ if $\beta_{(i)} \epsilon \left[L_j, U_j\right]$, otherwise $a_{ij} = 0$.

c. For matrix $\boldsymbol{A}$ obtain:

   Sum of columns = $SC_i = \sum_{j=1}^{N} a_{ij}$

   Sum of rows = $SR_j = \sum_{i=1}^{N} a_{ij}$

   $w = max(SC_i, SR_j)$ for all i, j

d. Select the subject $a_{(w)}$ associated with w; denote $\beta_{(w)}$ and (L_w, U_w) as the parameter and confidence interval associated to $a_{(w)}$.

e. If $\beta_{(i)} \in [L_w, U_w]$ then $a_{(i)}$ belongs to W, let N_w be the number of elements of W

   i. Define a submatrix of $\boldsymbol{A}$, named $\boldsymbol{A}_{(\boldsymbol{w})}$ (N_w x N_w), the columns and rows of $\boldsymbol{A}_{(\boldsymbol{w})}$ are subjects of W, and the elements are $a_{ij} = 1$ if $\beta_{(i)} \epsilon \left[L_j, U_j\right]$, otherwise $a_{ij} = 0$; $i = 1, \dots, N_w$, $j = 1, \dots, N_w$.

   ii. For $\boldsymbol{A}_{(\boldsymbol{w})}$ obtain:

   Sum of columns = $SC_{wi} = \sum_{j=1}^{N_w} a_{ij}$

   Sum of rows = $SR_{wj} = \sum_{i=1}^{N_w} a_{ij}$

   $v = min(SC_{wi}, SR_{wj})$ for all i, j

   iii. Select $a_{(v)}$ associated with v.

   iv. Let $q = 0.3$. If $v \geq (q * N_w)$ then assign the elements of W to a cluster G_w, otherwise remove $a_{(v)}$ from W (now $N_w = N_w - 1$) and repeat steps i) trough iv)

f. Remove subjects assigned to a cluster and with those that remain repeat steps 2b) to 2e)

g. Stop the process when all the subjects $(a_{(i)})$ have been assigned to one cluster (G_g).

3. Compare the obtained clusters, let $I^k = (I_{k1}, \dots, I_{kG})$ be the set of clusters associated with the k-th explanatory variable

   a. Sort subjects by $I^1, I^2, \dots, I^k$
   b. The final clusters are given by column k, see Figure 1
   c. Rename the clusters to a sequential order
4. Repeat step 1, if follow to step 2 increase the value of q by 0.1

Some groups may have only one subject, in that case, the subject is so different that it is necessary to analyze it separately. One advantage of this method is that the number of clusters is determined based on data variability, we do not need to define or estimate the number of groups in advance, as it happens in k-means or methods proposed by Su et al. [5] and Lin et al. [6]. The clustering algorithm was programmed in R software version 3.4.0 [13]; functions of plm package version 1.6-5 were used [10]. The program is available as supplementary material.

4 Classification Performance and Simulation

Su et al. [4] defined the following sequences of events to evaluate a classification method

$$\widehat{\boldsymbol{EI}}_{\boldsymbol{g},\boldsymbol{i}} = \{\boldsymbol{a}_{(i)} \notin \widehat{\boldsymbol{G}}_{\boldsymbol{g}} | \boldsymbol{a}_{(i)} \in \boldsymbol{G}_{\boldsymbol{g}}\} \text{ and } \widehat{\boldsymbol{EII}}_{\boldsymbol{g},\boldsymbol{i}} = \{\boldsymbol{a}_{(i)} \in \widehat{\boldsymbol{G}}_{\boldsymbol{g}} | \boldsymbol{a}_{(i)} \notin \boldsymbol{G}_{\boldsymbol{g}}\}$$

(11)

$$\widehat{\boldsymbol{EI}}_{\boldsymbol{g}} = \cup_{i \in G_g} \widehat{\boldsymbol{EI}}_{\boldsymbol{g},\boldsymbol{i}} \text{ and } \widehat{\boldsymbol{EII}}_{\boldsymbol{g}} = \cup_{i \in \widehat{G}_g} \widehat{\boldsymbol{EII}}_{\boldsymbol{g},\boldsymbol{i}} \qquad (12)$$

where $\widehat{EI}_{g,i}$ denotes the error event of not classifying an element $a_{(i)}$ of G_g into the estimated group $\widehat{\mathrm{G}}_g$, and $\widehat{EII}_{g,i}$ denotes the error event of classifying an element $a_{(i)}$ that does not belong to G_g into the estimated group $\widehat{\mathrm{G}}_g$. The events $\widehat{EI}_g$ and $\widehat{EII}_g$ mimic the Type I and Type II errors in statistical tests. To measure classification accuracy we use classification errors introduced by Su et al. [5]

$$\bar{P}(\widehat{EI}) = \frac{1}{N} \sum_{g=1}^{G} \hat{P}\left(\widehat{EI}_g\right) \text{ and } \bar{P}(\widehat{EII}) = \frac{1}{N} \sum_{g=1}^{G} \hat{P}\left(\widehat{EII}_g\right) \qquad (13)$$

where $\widehat{\mathrm{P}}$ denotes the relative frequency (empirical probability). The accuracy estimation of the parameters is evaluated with the root mean square error (RMSE), weighted by the proportion in the population ([6], [5])

$$RMSE = \left\{\frac{1}{MGKN}\sum_{j=1}^{M}\sum_{g=1}^{G}\sum_{k=1}^{K}N_g\left[\hat{\beta}_{gk,m}(\widehat{\mathrm{G}}_g)-\beta_{gk}(G_g)\right]^2\right\}^{1/2} \tag{14}$$

where $\hat{\beta}_{gk,m}(\hat{G})$ is the pooled slope parameter, estimated for the k-th variable and the g-th group $\widehat{\mathrm{G}}_g$ in the m-th replication, and $\beta_{gk}(G_g)$ is the true slope parameter for the k-th variable in the g-th group G_g. M is the number of replications in a Monte Carlo simulation study.

4.1 Simulation

The performance of the clustering methodology was examined using Monte Carlo simulations. The data were generated from the panel structure model (8), equivalently

$$y_{it}^0 = \sum_{g=1}^{G}\left[\sum_{i\epsilon G_g}\beta_{0i}^0 + \sum_{i\epsilon G_g}\sum_{k=1}^{K}x_{itk}^0\beta_{gk}^0\right] + \varepsilon_{it}^0 \tag{15}$$

The superscript zero indicates a known value. Four data generating process (DGP) were considered; M=1000 replications were generated for different combinations of subjects and time, N=(20,50,100,200), T=(10,20,50,100,200).

DGP1 a). Fixed model with 3 groups and 1 explanatory variable
(G, K) = (3,1); $(\beta_1^0, \beta_2^0, \beta_3^0) = (0.3, 0.5, 0.8)$; $(N_1^0, N_2^0, N_3^0) = (0.2N, 0.4N, 0.4N)$; $x_{it}^0 \sim N(1,3)$; $(\beta_{01}^0, \beta_{02}^0, \beta_{03}^0) = (1,5,8)$; $\varepsilon_i^0 \sim N(0,1)$.

DGP1 b). Same than DGP1 a), only fixed model was changed to random model with $\beta_{0i}^0 \sim N(1,1)$

DGP2. Random model with 3 groups and 1 explanatory variable
(G,K)=(3,1). $(\beta_1^0, \beta_2^0, \beta_3^0) = (0.5, 1.5, 2)$. $(N_1^0, N_2^0, N_3^0) = (0.2N, 0.4N, 0.4N)$. $x_{it}^0 \sim N(1,3)$. $\beta_{0i}^0 \sim N(1,1)$. $\varepsilon_i^0 \sim N(0,1)$.

DGP3. Fixed model with 3 groups and 2 explanatory variables

(G,K)=(3,2). $(\beta_{k1}^0, \beta_{k2}^0, \beta_{k3}^0) = \left(\begin{pmatrix}0.4\\1.6\end{pmatrix}, \begin{pmatrix}1\\1\end{pmatrix}, \begin{pmatrix}1.6\\0.4\end{pmatrix}\right)$.

$(N_1^0, N_2^0, N_3^0) = (0.2N, 0.4N, 0.4N)$. $x_{it1}^0 \sim N(1{,}3)$. $x_{it2}^0 \sim N(1.5{,}3)$.

$(\beta_{01}^0, \beta_{02}^0, \beta_{03}^0) = (1{,}5{,}8)$. $\varepsilon_i^0 \sim N(0{,}1)$.

DGP4. Random model with 4 groups and 3 explanatory variables

(G,K)=(4,3). $(\beta_{k1}^0, \beta_{k2}^0, \beta_{k3}^0, \beta_{k4}^0) = \left(\begin{pmatrix}0.4\\1.6\\2\end{pmatrix}, \begin{pmatrix}1\\1\\1\end{pmatrix}, \begin{pmatrix}1.6\\0.4\\3\end{pmatrix}, \begin{pmatrix}2\\1.5\\1\end{pmatrix}\right)$.

$(N_1^0, N_2^0, N_3^0, N_4^0) = (0.2N, 0.3N, 0.4N, 0.1N)$. $x_{it1}^0 \sim N(1{,}3)$. $x_{it2}^0 \sim N(1.5{,}3)$.

$x_{it3}^0 \sim N(2{,}3)$ $\beta_{0i}^0 \sim N(1{,}1)$. $\varepsilon_i^0 \sim N(0{,}1)$.

The clustering algorithm was executed with the values of $q = 0.3$ and $\alpha = 0.02$. Table 1 shows the classification errors and RMSE for GDP1 a) and GDP1 b), it can be seen that RMSE increases in the random model, this is to be expected because in random model there are two sources of variation, $\beta_{0i}^0 \sim \mathrm{N}(\mu_{\beta_{0i}^0}, \sigma_{\beta_{0i}^0}^2)$ and $\varepsilon_i^0 \sim \mathrm{N}(\mu_{\varepsilon_i^0}, \sigma_{\varepsilon_i^0}^2)$, if both are independent the variance increases to $\sigma_{\beta_{0i}^0}^2 + \sigma_{\varepsilon_i^0}^2$, in the analyzed case the variance doubles from 1 to 2.

Table 1: Estimations of $\bar{P}(\widehat{EI})$, $\bar{P}(\widehat{EII})$ and RMSE for different values of N and T. DGP1 a) and DGP1 b) are considered

		DGP1 a)			DGP1 b)		
N	T	$\bar{P}(\widehat{EI})$	$\bar{P}(\widehat{EII})$	RMSE	$\bar{P}(\widehat{EI})$	$\bar{P}(\widehat{EII})$	RMSE
20	10	0.22440	0.32695	0.161282	0.14510	0.46495	0.285527
20	20	0.20755	0.18865	0.117204	0.19420	0.36085	0.278231
20	50	0.10775	0.03250	0.178777	0.15455	0.18260	0.273337
20	100	0.05580	0.00105	0.130275	0.08220	0.04595	0.270784
50	10	0.24896	0.33994	0.080765	0.15472	0.47158	0.280031

50	20	0.23846	0.20558	0.048743	0.21056	0.37218	0.276321
50	50	0.12288	0.03586	0.016480	0.19244	0.20106	0.272203
50	100	0.05484	0.00144	0.005612	0.10570	0.05284	0.270304
100	10	0.25555	0.34558	0.079471	0.15843	0.47335	0.276097
100	20	0.25383	0.21261	0.046934	0.22187	0.36897	0.274984
100	50	0.13606	0.03855	0.014523	0.21518	0.21692	0.271609
100	100	0.05491	0.00195	0.004069	0.11956	0.05075	0.270396
200	10	0.26252	0.34301	0.079410	0.16055	0.47568	0.274443
200	20	0.26666	0.21946	0.047671	0.22577	0.36824	0.274844
200	50	0.15796	0.04454	0.014864	0.23802	0.22986	0.271466
200	100	0.05657	0.00290	0.002916	0.13638	0.05380	0.271405

Table 2: Estimations of $\bar{P}(\widehat{EI})$, $\bar{P}(\widehat{EII})$ and RMSE for different values of N and T. DGP2, DGP3 and DGP4 are considered

N	T	DGP2			DGP3			DGP4		
		$\bar{P}(\widehat{EI})$	$\bar{P}(\widehat{EII})$	RMSE	$\bar{P}(\widehat{EI})$	$\bar{P}(\widehat{EII})$	RMSE	$\bar{P}(\widehat{EI})$	$\bar{P}(\widehat{EII})$	RMSE
20	10	0.21655	0.22300	0.32021	0.32015	0.04040	0.71775	0.41050	0.01720	0.21185
20	20	0.20655	0.09950	0.31441	0.11670	0	0.32726	0.25435	0.00045	0.18102
20	50	0.07220	0.00045	0.30626	0.07300	0	0.25710	0.11095	0	0.16846
20	100	0.05320	0	0.30487	0.05710	0	0.23812	0.08090	0	0.16588
50	10	0.23744	0.20936	0.30927	0.33442	0.04332	0.06655	0.54856	0.01324	0.19123
50	20	0.24870	0.12874	0.30792	0.15958	0.00012	0.01913	0.36616	0.00058	0.17325
50	50	0.06554	0.00042	0.30417	0.06222	0	0.00518	0.15168	0	0.16672
50	100	0.05106	0	0.30302	0.05040	0	0.00365	0.11522	0	0.16296
100	10	0.25321	0.19457	0.30857	0.32591	0.03839	0.03337	0.63278	0.00941	0.18445
100	20	0.26127	0.13876	0.30655	0.20791	4.0E-05	0.01014	0.45898	0.00092	0.16913
100	50	0.06761	0.00051	0.30334	0.06205	0	0.00375	0.22183	0	0.16363
100	100	0.05196	0	0.30331	0.05116	0	0.00254	0.18330	0	0.16279
200	10	0.26409	0.18827	0.30598	0.29549	0.03226	0.02536	0.71229	0.00726	0.17552
200	20	0.28315	0.15016	0.30570	0.28094	7.0E-05	0.01080	0.57221	0.00111	0.16692
200	50	0.06874	0.00093	0.30315	0.06065	0	0.00275	0.32584	0	0.16335
200	100	0.05075	0	0.30293	0.05005	0	0.00186	0.28640	0	0.16273

Table 2 shows the classification errors and RMSE for GDP2, GDP3 and GDP4. As shown in Tables 1 and 2, if N increases the RMSE decreases; if the variance increases the classification errors increase. The values of $\bar{P}(\widehat{EI})$ and $\bar{P}(\widehat{EII})$ increase when the number of explanatory variables increases and decrease when T increases; $\bar{P}(\widehat{EI})$ and $\bar{P}(\widehat{EII})$ were less than 0.48 and when N and T were greater than 50 they were less than 0.2. Based on the information analyzed here, it can be inferred that the clustering classification method is acceptable for values of N and T greater than 50, but in some circumstances, it can be considered even for minor values.

5 Application to Trend Estimation of GDP Per Capita

One of the most important variables in economics is Gross Domestic Product (GDP) per capita as an indicator of a country's standard living. The data used in this study were obtained from the world development indicators of "The world Bank" web page [15]. Data of GDP per capita by country were analyzed by Su et al. [5], they found differences in GDP per capita growth between countries and proposed 4 clusters; using the same data and period that they used, the hypothesis test of slope homogeneity is rejected in all their clusters with p-values 7.59e-74, 7.66e-47, 2.29e-86 and 1.96e-46 for clusters 1,2,3 and 4 respectively, therefore, assumption 2 is not satisfied and we cannot apply panel data theory within each cluster, the previous results implies that the method proposed by Su et al. [5] may be improved. The method discussed here allows determining clusters that share a similar economic growth based on GDP and eliminates the arbitrary selection of groups. Annual data from 1960 to 2015 were analyzed; we used information of GDP (current US$), industry value (% of GDP) and agriculture value (% of GDP). Three models were analyzed, one that considers only GDP per capita by year in order to obtain clusters of countries that share the same economic growth; a second one, that considers GDP per capita as dependent variable and industry value as explanatory variable and a third one, that add agriculture value as a second explanatory variable .

5.1 Model 1

By deleting countries with many missing values in the period 1960 to 2015, an unbalanced panel data was obtained with N=96 countries and T=56 observations by country. Similar to Su et al. [5], we took logarithm and demean the data for each country as

$$y_{it} = ln(x_{it}) - \frac{1}{t}\sum_{1}^{t} ln(x_{it}) \quad i = 1, \dots ,96; \quad t = 1, \dots ,56 \tag{16}$$

where x_{it} is the GDP per capita by country. Model (8) was used with, y_{it} defined in equation (16) and x_{it} denoted the year. The clustering algorithm program was executed with $q = 0.3$ and $\alpha = 0.02$, as a result, 15 clusters were obtained; the cluster trend (CT) was estimated as the mean value of countries that belong to the same cluster.

$$CT_g = \frac{1}{n_g}\sum_{i \in G_g} y_{it} \qquad t = 1, \dots , \mathrm{T}; \quad g = 1, \dots , G \tag{17}$$

Table 3 shows names of countries included in each cluster and Figure 2 shows cluster trend of GDP per capita. Next some comments are given; clusters 1, 2, 3 and 4 contains 66% of the countries; two countries made a cluster by itself, Ireland and Korea Rep.; the countries of clusters 9, 13 and 14 (Botswana, Singapore, Ireland, Korea) have the highest growth of GDP per capita in the period; clusters 10 and 12 (Madagascar, Niger, Zimbabwe, Malaysia, Liberia, Luxembourg) started with the highest GDP per capita and finished with the lowest GDP per capita. The classification showed in Table 3 differs from the classifications of Su et al. [5]. A classification based on external criteria such as continental location or OECD countries is misleading.

We notice that almost all the clusters have a high growth in the period of 1960 to 1980, after 1980 the growth tendency change, because of that, the GDP per capita trend was analyzed for 1980 to 2015 (T=36), under this scenario 10 clusters were obtained, Table 4 shows the countries included in each cluster and Figure 3 shows the cluster trend.

Table 3: Classification of 96 economies based on GDP per capita for period 1960-2015

Cluster	Economies	Countries
1	19	Algeria, Bahamas, Belize, Canada, Congo Rep., Ecuador, Fiji, Gabon, Iran Islamic Rep., Morocco, Nigeria, Panama, Peru, Rwanda, Suriname, Sweden, Trinidad and Tobago, United States, Uruguay
2	17	Argentina, Bangladesh, Benin, Bolivia, Burkina Faso, Cameroon, Chad, Guyana, Honduras, Kenya, Malawi, Mauritania, Nepal, Nicaragua, Papua New Guinea, Sudan, Venezuela RB
3	17	Australia, Belgium, Brazil, Colombia, Denmark, Dominican Republic, Finland, France, Greece, Iceland, Israel, Italy, Lesotho, Mexico, Netherlands, Turkey, United Kingdom
4	11	Austria, Bermuda, China, Hong Kong SAR China, Japan, Norway, Portugal, Seychelles, Spain, St. Kitts and Nevis, St. Vincent and the Grenadines
5	5	Burundi, Central African Republic, Senegal, Sierra Leone, Zambia
6	6	Guatemala, India, Jamaica, Pakistan, Philippines, South Africa
7	4	Chile, Costa Rica, Sri Lanka, Swaziland
8	4	Cote dIvoire, Ghana, Togo, Uganda
9	2	Botswana, Singapore
10	3	Madagascar, Niger, Zimbabwe, Malaysia
11	2	Puerto Rico, Congo Dem. Rep.
12	2	Liberia, Luxembourg
13	2	Luxembourg, Thailand
14	1	Ireland
15	1	Korea Rep.

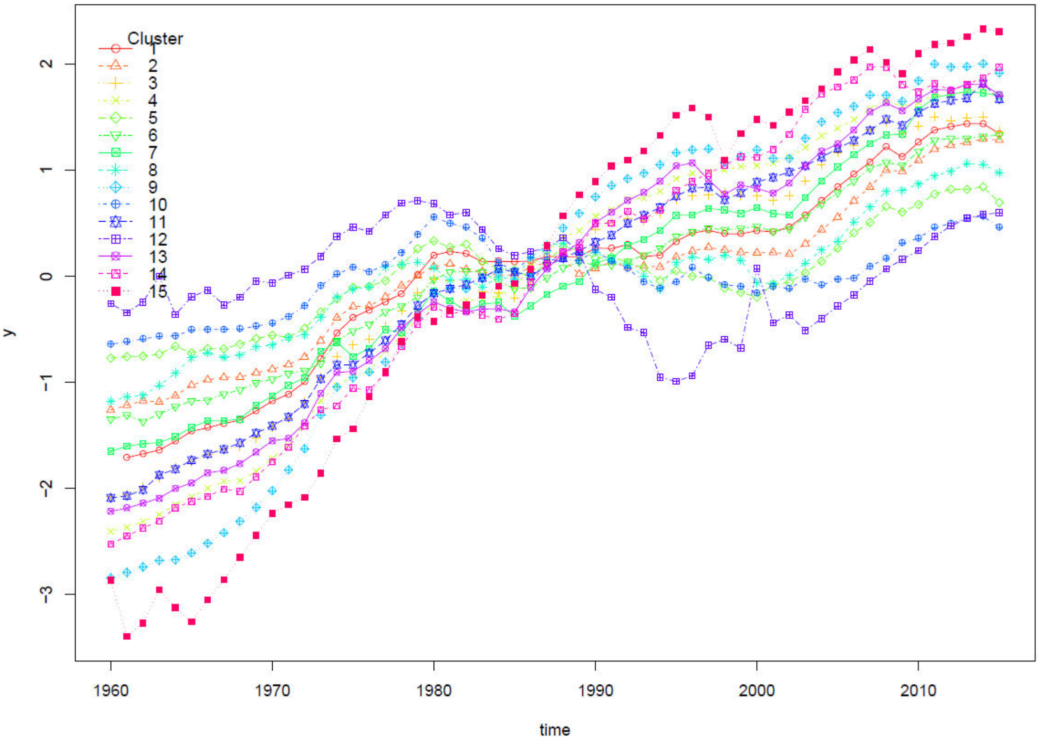

Figure 2: Cluster trend for period 1960-2015

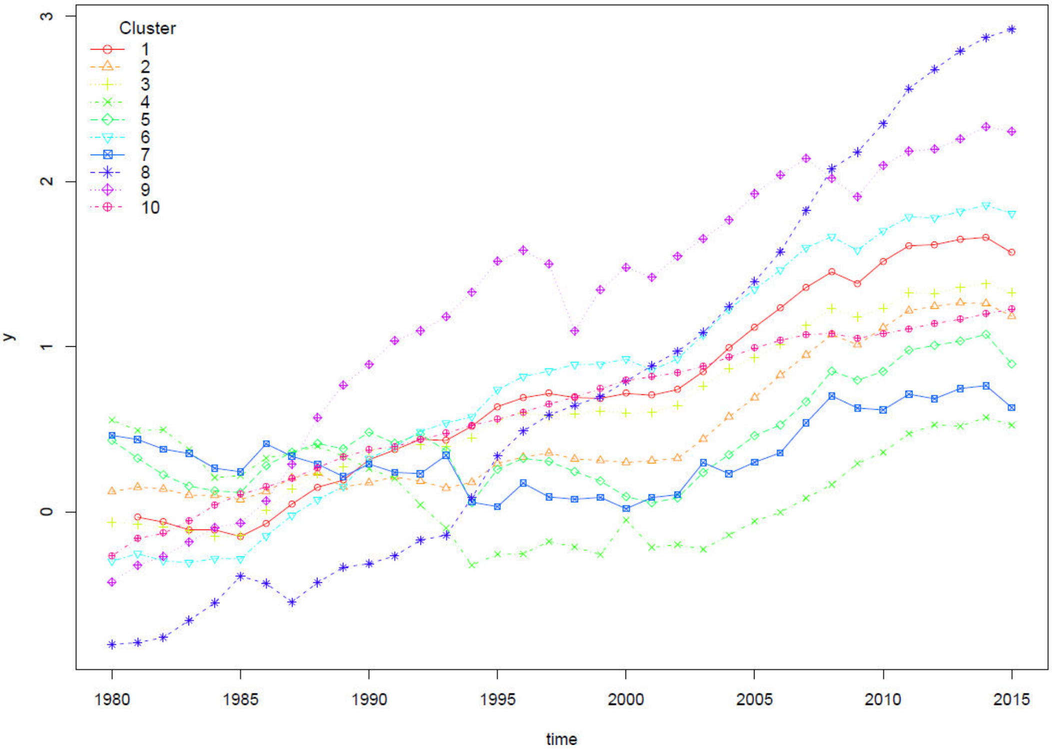

Figure 3: Cluster trend for period 1980-2015

Table 4: Cluster classifications of 96 economies based on GDP per capita for period 1980-2015

Cluster	Economies	Name of countries
1	31	Australia, Austria, Botswana, Brazil, Chad, Colombia, Costa Rica, Dominican Republic, Greece, Guyana, Hong Kong SAR China, India, Israel, Lesotho, Malaysia, Mexico, Nigeria, Norway, Peru, Portugal, Puerto Rico, Seychelles, Spain, St. Kitts and Nevis, St. Vincent and the Grenadines, Suriname, Swaziland, Thailand, Trinidad and Tobago, United Kingdom, Uruguay
2	27	Algeria, Argentina, Bahamas, Benin, Bolivia, Burkina Faso, Congo Rep., Ecuador, Fiji, France, Ghana, Guatemala, Honduras, Iceland, Iran Islamic Rep., Japan, Kenya, Malawi, Mauritania, Nicaragua, Sierra Leone, South Africa, Sudan, Sweden, Uganda, Venezuela RB, Zambia
3	14	Bangladesh, Belgium, Belize, Canada, Denmark, Finland, Italy, Jamaica, Morocco, Nepal, Netherlands, Pakistan, Panama, Philippines
4	6	Burundi, Central African Republic, Congo Dem. Rep., Liberia, Niger, Zimbabwe
5	6	Cote dIvoire, Gabon, Papua New Guinea, Rwanda, Senegal, Togo
6	7	Bermuda, Chile, Ireland, Luxembourg, Singapore, Sri Lanka, Turkey
7	2	Cameroon, Madagascar
8	1	China
9	1	Korea Rep.
10	1	United States

As can be seen from Figure 3 and Table 4, three countries made a cluster by itself, China, Korea Rep. and Unites States; China and Korea Rep. (clusters 8 and 9) are the countries with the lowest GDP per capita at the beginning of the series and the highest GDP per capita at the end of the series; all the countries in Clusters 4, 5 and 7 are African countries and have minimal grown in GDP per capita; cluster 2 increases the GDP per capita after the year 2000; clusters 1, 2 and 3 contains 75% of the countries, they have a regular GDP per capita growth; except for African countries, the cluster classification does not exhibit geographic features.

Table 5 reports p-values for different tests and estimations of parameters for each cluster. In all clusters, the hypothesis (5) is rejected which implies that exists panel effect, meanwhile, assumption 3 is satisfied, so there is no difference in the slope growth within each cluster and we can apply panel data theory; finally, the hypothesis (6) is not rejected, therefore, random model effects is preferred in all the clusters. The parameters were estimated with the random effects model; estimations of component errors are not shown here because by now, they are not part our objective. As can be seen from Table 5, cluster 8 has the highest slope, for cluster 4 the hypothesis of slope equal to zero is not rejected (p-value=0.5847705), so cluster 4 has not GDP growth in the analyzed period.

Table 5: Test of hypothesis (p-value) and estimation of parameters for model 1

				Estimation of parameters			
Cluster	Test (3)[a]	Test (5)[b]	Test (6)[c]	β_0	*p-value* (β_0)	β_1	*p-value* (β_1)
1	0.17817	5.47E-74**	0.84086	-111.4612	0**	0.05615	0**
2	0.11512	2.55E-42**	0.90855	-71.2221	6.3E-193**	0.03590	1.1E-194**
3	0.55144	3.12E-45**	1	-91.6811	7.4E-244**	0.04620	3.3E-245**
4	0.56142	7.72E-06**	1	2.8941	0.5648091	-0.00137	0.5847705
5	0.43134	6.07E-10**	0.97431	-38.9763	1.3E-23**	0.01973	5.2E-24**
6	0.07625	1.16E-31**	0.92143	-138.9611	3.1E-149**	0.06996	7.8E-150**
7				-19.5764	0.004285**	0.00997	0.00368**
8				-225.1367	9.54E-26**	0.11307	8.5E-26**
9				-155.5692	2.35E-20**	0.07849	1.8E-20**
10				-81.5733	4.90E-29**	0.04114	3.8E-29**

[a] H_0: Same coefficients apply to each subject

[b] H_0: OLS is better than random effects model, i.e. no panel effect

[c] H_0: Preferred model is random effects vs. H_1: preferred model is fixed effects

* significant at 5%; ** significant at 1%

4.2 Model 2

Now, the dependent variable is GDP per capita and industry value in % of GDP (x_1) is the explanatory variable, the model is the following

$$y_{it} = \beta_{0i} + x'_{1it}\beta_1 + \varepsilon_{it} \tag{18}$$

where $y_{it} = ln(GDP)$, $x_{1it} = ln(x_1)$ and ε_{it} is the error term. After removing countries that do not have industry information for the series of 1980 to 2015, we have a balanced panel data, with N=61 countries and T=36 observations by country. If all countries are pooled, the hypothesis of parameter homogeneity (3), is rejected with a *p-value = 2.2 e-16*, therefore we cannot use panel data theory for the whole group of countries. We executed the clustering algorithm with values of $q = 0.55$ and $\alpha = 0.02$. Table 6 reports the cluster classification, 13 clusters were obtained.

Table 6: Cluster classifications of 61 economies based on the relation of GDP per capita and industry value for the period 1980-2015 (Model 2)

Cluster	Economies	Countries
1	6	Bangladesh, China, India, Kenya, St. Kitts and Nevis, Thailand
2	11	Australia, Belize, Botswana, Brazil, Cameroon, Chile, Fiji, Finland, Guyana, St. Vincent and the Grenadines, Turkey
3	15	Algeria, Burkina Faso, Burundi, Central African Republic, Chad, Colombia, Dominican Republic, Honduras, Iran Islamic Rep., Malaysia, Nepal, Senegal, Togo, Venezuela RB, Zambia
4	6	Bolivia, Ecuador, Mexico, Norway, Panama, Suriname
5	6	Denmark, Morocco, Pakistan, Philippines, Singapore, Sweden
6	6	Congo Dem. Rep., Congo Rep., Lesotho, Madagascar, Mauritania, Uganda
7	3	Argentina, France, Netherlands
8	2	Benin, Liberia
9	2	Malawi, South Africa
10 – 13	1 by cluster	Puerto Rico (10), Austria (11), Korea Rep. (12), Sierra Leone (13)

Table 7: Test of hypothesis (p-value) and estimation of parameters for model 2

				Estimation of parameters			
Cluster	Test (3)[a]	Test (4, 5)[b]	Test (6)[c]	β_0	*p-value*(β_0)	β_1	*p-value*(β_1)
1	0.56824	2.9E-61**	0.01030*			4.7095	1.2E-21**
2	0.71989	1.2E-136**	0.00038**			-2.2892	6.1E-31**
3	0.96419	7.0E-90**	2.9E-16**			-0.0504	0.73367
4	0.98558	6.7E-80**	0.86334	0.5070	0.59793	2.2999	6.6E-19**
5	0.54753	1.3E-139**	0.19561	26.248	4.7E-46**	-5.2681	2.2E-36**
6	0.10341	1.7E-26**	0.95456	2.8380	9.8E-18**	1.0200	2.0E-28**
7	0.39599	4.4E-23**	0.40220	20.836	6.4E-65**	-3.3838	1.8E-43**
8				4.0763	3.9E-27**	0.6303	2.2E-10**
9				-42.980	0.00178**	14.409	0.00023**
10				-3.8345	0.00163**	3.2647	1.2E-13**
11				34.648	1.0E-2**	-7.0833	4.5E-17**
12				-29.676	4.1E-08**	8.7854	9.6E-11**
13				6.9045	1.1E-20**	-0.4348	0.00088**

[a] H_0: Same coefficients apply to each subject

[b] H_0: OLS are better than fixed effects model (random effects model),

[c] H_0: Preferred model is random effects vs H_1: Preferred model is fixed effects

* significant at 5%; ** significant at 1%

Table 7 reports the p-value for different tests and the estimation of parameters for each cluster, it can be noticed that all clusters have panel effect and the same coefficient apply within each cluster. Fixed effect model is preferred for clusters 1, 2 and 3 and random effects model is preferred for the rest of clusters, estimations of β_{0i} values according to the preferred model are in Table 8.

Table 8: Values of β_0 for clusters where the preferred model is fixed effects (Model 2)

Clusters 1		Clusters 2		Clusters 3	
Country	β_0	Country	β_0	Country	β_0
Bangladesh	-8.86577	Australia	17.84477	Algeria	8.04547
China	-11.1085	Belize	14.96266	Burkina Faso	5.94919
India	-9.33262	Botswana	16.86114	Burundi	5.34768
Kenya	-7.64363	Brazil	16.24878	C. African Republic	5.98960
St. Kitts and Nevis	-6.41712	Cameroon	14.64212	Chad	5.95519
Thailand	-9.21436	Chile	16.75221	Colombia	7.97443
		Fiji	14.87301	Dominican Rep.	7.90874
		Finland	18.16704	Honduras	7.15881
		Guyana	14.55196	Iran Islamic Rep.	8.15994
		St. Vincent and the Grenadines	14.83356	Malaysia	8.48464
		Turkey	15.96111	Nepal	5.71672
				Senegal	6.67851
				Togo	6.06009
				Venezuela RB	8.61715
				Zambia	6.55740

Based on Tables 6, 7 and 8, some comments are given; 4 clusters have only one country; cluster 3 is the cluster with more countries (15) and the hypothesis $H_0: \beta_1 = 0$ is not rejected with a *p-value*= 0.733677 so it can be inferred that the GDP per capita of the countries in cluster 3 has no relation with industry; all the countries of clusters 6, 8 and 9 are from Africa; the clusters with a positive relation between industry and GDP are 1, 4, 6, 8, 9, 10 and 12, meanwhile the clusters with negative relations are 2, 5, 7, 11 and 13; clusters 1, 9 and 12 are the top three clusters with positive relation between industry and GDP, they contain countries with a high industry level like Bangladesh, China, India, Thailand, South Africa and Korea Rep.; cluster 11 (Austria) is the one with the highest negative slope, i.e. GDP is negatively influenced by industry, but we need to consider that it has the highest intercept value ($\beta_{0(11)}$).

4.3 Model 3

As an extension to multiple variables, a new explanatory variable is added to model 2, the model is the following

$$y_{it} = \beta_{0i} + x'_{1it}\beta_1 + x'_{2it}\beta_2 + \varepsilon_{it} \tag{19}$$

where x_2 is agriculture value in % of GDP and $x_{2it} = ln(x_2)$, the other variables were defined in model 2. In order to use panel theory it is necessary that the parameters associated with the explanatory variables follow a similar trend, again hypothesis of parameter homogeneity (3) is rejected with a *p-value*=0.002; after applying clustering algorithm with $q = 0.55$ and $\alpha = 0.02$, 30 clusters were obtained, 13 clusters have only one country, 5 clusters have 2 countries and 6 clusters have three or more countries, the economies by cluster are showed in Table 9.

Table 10 reports *p*-values for different tests and Table 11 reports estimation of parameters for each cluster. The p-value of parameters β_1 and β_2 is not significant at 0.05 for clusters 12 and 22 (Burkina Faso y Uganda) so for this two countries the GDP per capita has not relation with industry and agriculture; the p-value for the explanatory variable industry in clusters 5, 13 and 21 (Bolivia, India, Kenya, Mexico, Fiji) is not significant at 0.05 so if the variable agriculture is present, the variable industry has no relation with the GDP per capita for those countries. The p-value of the explanatory variable agriculture is not significant at 0.05 in clusters 11, 19, 23, 25 and 27 (Central African Republic, Togo, Argentina (19), Thailand (23), Sierra Leone (25), Madagascar (27)) so if the variable industry is present, the variable agriculture has no relation with the GDP per capita for those countries. As we can figure out, if the number of explanatory variables increases, then the number of clusters increases.

Table 9: Classification of 61 economies based on the relation of GDP per capita with industry and agriculture for 1980-2015 (Model 3)

Cluster	Economies	Countries
1	12	Australia, Bangladesh, Congo Dem. Rep., Dominican Republic, Ecuador, France, Guyana, Senegal, South Africa, St. Vincent and the Grenadines, Venezuela RB, Zambia
2	7	Belize, Chile, Colombia, Honduras, Norway, Panama, Turkey
3	4	Botswana, Denmark, Finland, Sweden
4	3	Benin, Brazil, Congo Rep.
5	3	Bolivia, India, Kenya
6	3	Cameroon, Iran Islamic Rep., Philippines
7	2	Burundi, Malawi
8	2	Mauritania, Suriname
9	2	Austria, Netherlands
10	2	Lesotho, St. Kitts and Nevis
11	2	Central African Republic, Togo
12 - 30	1 by cluster	Burkina Faso (12), Mexico(13), Puerto Rico (14), Chad (15), Singapore (16), Algeria (17), Malaysia (18), Argentina (19), China (20), Fiji (21), Uganda (22), Thailand (23), Korea Rep. (24), Sierra Leone (25), Liberia (26), Madagascar (27), Morocco (28), Nepal (29), Pakistan (30)

Table 10: Test of hypothesis (p-value) for model 3

Cluster	Test (7)[a]	Test (10)[b]
1	0.09817	0.52276
2	0.58281	0.69080
3	0.87614	0.01651*

[a] H_0: Same coefficients apply to each subject

[b] H_0: Preferred model is random effects vs H_1: Preferred model is fixed effects

* significant at 5%; ** significant at 1%

Table 11: Estimation of parameters for model 3

	Estimation of parameters					
Cluster	β_0	*p-value*	β_1	*p-value*	β_2	*p-value*
1	12.08391	3.5E-101**	-0.37759	0.000525**	-1.35392	3.27E-90**
2	14.77670	5.49E-75**	-0.81715	5.47E-08**	-1.69734	1.6E-102**
3			-1.10238	2.19E-09**	-0.89355	1.73E-46**
4	14.14184	2.90E-46**	-0.87407	1.73E-11**	-1.60863	4.50E-37**
5	13.67674	7.89E-15**	-0.40152	0.144584	-1.88265	7.10E-14**
6	17.46628	8.53E-24**	-1.60497	1.52E-06**	-1.64582	7.60E-29**
7	9.60011	1.02E-14**	-0.51573	0.014446*	-0.74023	0.000135**
8	6.64059	6.10E-09**	1.16311	3.17E-05**	-1.18637	6.22E-18**
9	15.85886	3.01E-40**	-1.42505	4.32E-12**	-0.98256	1.34E-23**
10	13.94639	1.83E-28**	-1.04689	6.93E-06**	-1.76099	1.26E-35**
11	2.84921	0.058313	0.54499	0.028408*	0.38920	0.098946
12	-0.50651	0.890208	0.59452	0.386206	1.28046	0.064622
13	11.59111	5.68E-08**	-0.28052	0.546611	-1.30411	3.23E-17**
14	-4.98409	0.115613	3.26980	4.47E-05**	-0.67203	1.04E-11**
15	-12.3826	1.02E-07**	1.76176	1.66E-07**	3.66158	3.58E-12**
16	2.60800	0.406900	1.77770	0.047942*	-0.52472	6.59E-11**
17	25.55890	0.000113**	-3.09533	0.009273**	-2.38193	0.001653**
18	25.85305	5.05E-16**	-3.49820	3.97E-09**	-1.78501	6.67E-18**
19	20.45586	3.66E-18**	-3.47886	4.30E-10**	0.17431	0.481905
20	24.38955	2.34E-06**	-2.60894	0.023702*	-2.71344	2.43E-23**
21	11.32091	5.12E-13**	0.18348	0.647483	-1.44280	4.84E-09**
22	6.72153	0.025741*	0.34529	0.410204	-0.54217	0.276332
23	-8.61652	0.172071	4.89477	0.002149**	-0.51521	0.238872
24	-3.42041	0.276453	4.03073	3.61E-05**	-1.21288	6.75E-24**
25	6.80751	0.000474**	-0.43187	0.002464**	0.02300	0.955398
26	14.76399	1.11E-09**	-0.34094	0.039026*	-2.15872	9.86E-07**
27	4.21122	0.003055**	1.06799	0.000206**	-0.38230	0.236974
28	32.37882	4.59E-14**	-6.38840	6.70E-11**	-1.21880	0.000693**
29	20.31894	3.25E-18**	-1.48600	7.18E-08**	-2.80346	1.47E-15**
30	30.60846	1.19E-15**	-4.06140	2.31E-09**	-3.54512	1.35E-09**

* significant at 5%; ** significant at 1%

6 Conclusions

We proposed a method to determine clusters of subjects in panel data; the method is useful when exits parameter heterogeneity and standard panel data theory cannot be applied. An algorithm is proposed to cluster units into several groups such that within each cluster units share the value of the coefficients. The clustering is achieved by checking whether confidence intervals from different units overlap or not; the final clusters must have parameter homogeneity. Simulations were conducted, for values of N and T greater than 50 the clustering method had acceptable performance. The clustering method was carried out to study the heterogeneous trending behavior of GDP per capita across 96 countries for two periods, 1960-2015 and 1980-2015; 15 and 10 clusters were found respectively. Our preliminary investigation suggests that if the number of explanatory variables increases then the number of clusters increases.

References

[1] S. Durlauf, A. Kourtellos and A. Minkin, "“The Local Solow Growth Model”," *European Economic Review,* vol. 45, pp. 928-940, 2001.

[2] L. Su and Q. Chen, "Testing Homogeneity in Panel Data Models with Interactive Fixed Effects," *Econometric Theory,* vol. 29, pp. 1079-1135, 2013.

[3] J. Blomquist and J. Westerlund, "Testing slope homogeneity in large panel with serial correlation," *Economics Letters,* vol. 121, no. 3, pp. 374-378, 2013.

[4] L. Su, Z. Shi and P. Phillips, "Identifying Latent Structures in Panel Data," *Econometrica,* vol. 84, no. 6, pp. 2215-2264. http://ink.library.smu.edu.sg/cgi/viewcontent.cgi?article=2910&context=soe_research, 2016.

[5] L. Su, X. Wang and S. Jin, "Sieve Estimation of Time-Varying Panel Data Models with Latent Structures," *Research Collection School of Economics,* pp. 1-44. http://ink.library.smu.edu.sg/soe_research/1723, 2015.

[6] C. Lin and S. Ng, "Estimation of Panel Data Models with Parameter Heterogeneity when Group Membership is Unknown," *Journal of Econometric Methods,* vol. 1, no. 1, p. 42–55, 2012.

[7] A. Hoogstrate, F. Palm and G. Pfann, "Pooling in dynamic panel data models: An application to forecasting GDP growth rates," *Journal of Business & Economic Statistics,* pp. 274-283. http://digitalarchive.maastrichtuniversity.nl/fedora/get/guid:405d3649-06c5-44ae-8c27-a50a28717d47/ASSET1, 2000.

[8] S. Frühwirth-Schnatter and S. Kaufmann, "Model-based clustering of

multiple time series," *Journal of Business and Economic Statistics,* no. 26, pp. 78-89, 2008.

[9] R Core Team, "R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria," 2017. [Online]. Available: https://www.R-project.org/.

[10] Y. Croissant and G. Millo, "Panel Data Econometrics in R: The plm package," *Journal of statistical software,* pp. 1-43. https://www.jstatsoft.org/article/view/v027i02/v27i02.pdf, 2008.

[11] B. Baltagi, Econometric Analysis of Panel Data, 3rd ed., Hoboken, NJ: John Wiley & Sons, 2005.

[12] E. Fress, Longitudinal and Panel Data: Analysis and Applications in the Social Sciences, Cambridge University Press, 2004.

[13] W. H. Greene, Econometric analysis, 6th ed., Upper Saddle River, N.J.: Prentice Hall, 2008.

[14] J. Hausman, "Specification Tests in Econometrics," *Econometrica,* vol. 46, pp. 1251-1271, 1978.

[15] The World Bank, "DataBank, World development indicators," [Online]. Available: http://databank.worldbank.org/data/reports.aspx?Code=FP.CPI.TOTL.ZG&id=1ff4a498&report_name=Popular-Indicators&populartype=series&ispopular=y#. [Accessed 07 06 2017].