ACTAS DEL XV CONGRESO
DR. ANTONIO A. R. MONTEIRO (2019)

LONGITUDINAL ANALYSIS METHODOLOGIES TO DETERMINE PROFILES OF INFORMAL WORKERS FROM GRAN CÓRDOBA

MAXIMILIANO LUJÁN IGLESIAS AND MARÍA INÉS STIMOLO

ABSTRACT. One of the main limitations for the correct analysis of the labor market in developing countries from a temporal variability approach is the lack of appropriate panel data information. The purpose of this article is to develop statistical methodologies that make it possible to incorporate the temporal dimension as a key factor to carry out the most complete possible analysis of the dynamics and structure of the problem under study, as well as the relationship between its multiple factors and determinants.

The pseudo-panel approach allows overcoming the limitation of data availability through the building of synthetic panels and measuring the temporal evolution of characteristics of interest in cohorts of individuals and the building of "variables-trajectories".

The temporal clustering approach, based on the non-parametric k-means algorithm, combines content similarities and temporal adjacency in a single representation, which makes it possible to find cohort groups of homogeneous individuals in relation to the joint trajectories of their characteristics.

With the results obtained according to the exploratory analysis, we could identify three well-defined groups based on the temporal trajectories of their informality rates and income levels.

1. INTRODUCTION

It is currently estimated that more than sixty percent of the world's employed population are in the informal economy. This involves all economic activities performed by workers or economic units that are, in law or in practice, not covered or insufficiently covered by formal arrangements [17].

The conditions of vulnerability to which informal workers are exposed have brought to light the importance of studying and monitoring their evolution over time. However, a limitation for the correct analysis of the labor market in developing countries from a temporal variability approach is the lack of appropriate panel data information [5]. In some of these countries, panel data information from official programs is available but there are certain limitations. In surveys based on rotating panels, households or individuals remain a relatively short time period in the sample, which makes it impossible to follow them up over time. In addition, the non-random withdrawal of certain units due to "attrition" can produce a considerable bias in the estimations [21]. Other countries do not have panel household surveys but they rely on a series of independent cross sections from statistical programs conducted over relatively long time periods.

The pseudo-panel methodology allows overcoming the limitation of data availability through the building of synthetic panels. This is achieved by replacing the individual observations of the original panel by means of subgroups of the population, that is, subgroups of individuals whose appearance can be identified in repeated cross-sectional surveys [18]. This approach allows us to follow cohorts over time in repeated cross sections, producing time series for the means of the subgroups that can be used as if they were panel data, to estimate their temporal evolution and the building of "variables-trajectories".

This article aims at identifying groups with different characteristics that remain in time. Standard cluster techniques allow us to identify groups without taking into account the temporal aspect. Cluster-longitudinal techniques combine content similarities and temporal adjacency in a single representation. This implies that temporal clusters that take into account temporal "neighbors" of the objects must be used to extract useful knowledge for the most complete possible analysis of the dynamics and structure of the problem under study, as well as the relationship between its multiple factors and determinants.

It is possible to classify the cohorts using the longitudinal clustering technique that uses the KmL approach (K-means longitudinal) of Genolini and Falissard [11]. With this method, we can identify the joint evolution of homogeneous trajectories in the cohorts, enabling the formation of groups of workers in the analyzed time period, and the trend changes, possible temporal factors (alterations, technical changes), that affect each specific group.

Given that cohorts C are considered in the pseudo-panel analysis, a matrix $Y_{c..}$ of joint trajectories of dimension $P \times T$ is built for each cohort c, where P represents the estimated characteristics or attributes in each specific cohort at T different times.

2. Pseudo-panel model

2.1. **Standard linear regression.** To introduce the pseudo-panel approach, it is necessary to start by presenting the standard or typical linear regression model for cross-sectional data. The multiple linear regression model for a set of subjects N and variables P can be expressed in matrix form as

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

where y is an $N \times 1$ vector of the dependent variable of interest; X is a matrix of explanatory variables (or regressors) with dimension $N \times P$; β is the vector of parameters that indicates the effect of the explanatory variables on the dependent; and ε is a $P \times 1$ vector with stochastic disturbance terms.

In practice, we work with population samples, so the values of the vector β indicated in (1) are unknown and the random disturbance vector ε is unobservable.

The vector of parameters β is estimated by the ordinary least squares method that minimizes the sum of the squares of the remainders[1], resulting in the following vector of least quadratic estimators:

$$\hat{\beta} = \left[X'X\right]^{-1} X'y.$$

However, when there are different measurements in the subjects' time, an autocorrelation is used among the observations, which does not take into account this estimation since it is based on the limitation that observations are independent of each other. This resulted in the development of panel data models.

2.2. **Panel model.** In statistics and econometrics, the term *panel data* refers to data that combines a temporal dimension with a transversal dimension. That is, a set of individuals that are observed at different times.

The monitoring of observations over time provides information that allows a better study of the dynamics of change and a better explanation of the phenomena.

The typical panel model (2) adds an individual effect to the standard linear model to capture the effect of each individual on the time-dependent variable. This can be expressed

[1] $\min\limits_{\beta} \sum\limits_{i=1}^{N} \hat{\varepsilon}_i^2 = \min\limits_{\beta} \sum\limits_{i=1}^{N} (y - X\hat{\beta})(y - X\hat{\beta})$.

by the equation

$$\begin{aligned} y_{it} &= x_{it}\beta + \alpha_i + \varepsilon_{it}, \\ i &= 1,\ldots,N. \\ t &= 1,\ldots,T. \end{aligned}$$

where y_{it} is the variable of interest for the i-th observation at time t; x_{it} is the (linear) vector of the explanatory variables p; β is the vector of parameters; α_i is the individual effect that captures all the determinants of the variable of interest that are fixed in time; and ε_{it} is a disturbance term. The subscript t indicates the t-th time.

In a compact form, the model can be presented as follows:

$$\begin{aligned} y &= X\beta + C\alpha + \varepsilon \\ C &= I_N \otimes \iota, \end{aligned} \tag{2}$$

where y has dimension $NT \times 1$; X is $NT \times P$; β and ι are $T \times 1$ vectors; so C is $NT \times 1$; and ε is $N \times T$.

The first issue to determine is whether the unobservable random variable α_i is correlated with the vector of regressors x_{it} or if it is independent of the latter, on which the algorithm of the estimation of parameters depends.

In the first case, when there is correlation, it is convenient to perform the estimation under conditional inference, called the *fixed effects* panel model. This model has fewer assumptions with respect to errors and is usually the most consistent one. In that case, α_i assumes a fixed value for each individual and is estimated in conjunction with β by the ordinary least squares method (OLS). According to Arellano [3], we have, as a result, the following fixed effects estimator of the explanatory variables, also called *intra-groups* estimator:

$$\hat{\beta}_{FE} = \left(X'\bar{Q}X\right)^{-1}X'\bar{Q}y = \left(\tilde{X}'\tilde{X}\right)^{-1}\tilde{X}'\tilde{y},$$

where $\bar{Q} = I_{NT} - C(C'C)^{-1}C'$ has dimension $NT \times NT$, with $C = I_N \otimes \iota$ of dimension $NT \times 1$; $\tilde{X} = \bar{Q}X$ is $NT \times P$; and $\tilde{y} = \bar{Q}y$ is $NT \times 1$.

In the second case, when it is assumed that α_i is independent of x_{it}, it is convenient to perform the estimation under unconditional inference, called the *random effects* panel model. In that case, α_i is not a fixed value but a random component that is part of the disturbance term. Therefore, this model contains a compound disturbance term $u_{it} = \alpha_i + \varepsilon_{it}$. The estimation of the random effects models is made using the generalized least squares method (GLS), resulting in the following fixed effect estimator of the explanatory or *intra-groups* variables:

$$\hat{\beta}_{GLS} = \left(X'\left[I_N \otimes \hat{\Omega}^{-1}\right]X\right)^{-1}X'\left[I_N \otimes \hat{\Omega}^{-1}\right]y,$$

where Ω is the covariance matrix and its elements are calculated by

$$\hat{\omega}_{ts} = \frac{1}{N}\sum_{i=1}^{N}\hat{u}_{it}\hat{u}_{is}.$$

In practice, to determine whether the individual effect and the observed regressors are correlated and, therefore, to choose the most correct model, the Hausman specification test [16] is generally used. This statistical hypothesis test allows us to evaluate whether an estimator $\hat{\beta}_e$ that is more efficient is also consistent if compared to another alternative estimator $\hat{\beta}_c$ that is known to be consistent. This is achieved by evaluating whether the differences between the estimations of both models are systematic or not. In case there are no systematic differences, both estimators would be consistent and, therefore, $\hat{\beta}_e$ would be a better estimator which is also more efficient.

Considering the compliance with the intrinsic assumptions of each model, the *fixed effects* estimator $\hat{\beta}_{FE}$ is consistent, while the *random effects* estimator calculated by generalized least squares $\hat{\beta}_{GLS}$ is more efficient (asymptotically) but inconsistent when the model is poorly specified. That is,

$$\text{Var}\left(\hat{\beta}_{GLS}\right) \leq \text{Var}\left(\hat{\beta}_{FE}\right).$$

The Hausman test is then defined as

$$h = \left[\hat{\beta}_{FE} - \hat{\beta}_{GLS}\right]' \left[\text{Var}\left(\hat{\beta}_{FE}\right) - \text{Var}\left(\hat{\beta}_{GLS}\right)\right]^{-1} \left[\hat{\beta}_{FE} - \hat{\beta}_{GLS}\right].$$

In the panel models, one of the limitations that affect the estimation is the availability and content of the data. That is, the presence of measurement errors due to the lack of answers or false answers in surveys (usual in income variables), rotating panels (sampling in household surveys) where the panel remains a relatively short time period in the sample, which makes it impossible to follow, and the effect of *attrition* or non-random withdrawal of individuals.

2.3. **Pseudo-panel model.** The pseudo-panel approach was initially presented by Nobel Prize winner Angus Deaton in 1985 with the aim of overcoming these last mentioned limitations by building *synthetic* panels [7]. This is achieved by replacing the individual observations of the original panel by means of subgroups of the population whose appearance can be identified in repeated cross-sectional surveys [18].

In order to define the subgroups, variables-factors which must be considered invariant over time (for example: year of birth, gender, ethnicity) must be selected. There are some assumptions that must be taken into account at the time of their building. In general terms, we must ensure that the cohorts are built on a stable population and criteria that ensure that their profiles do not change abruptly over time. From the selection of the variables-factors, those individuals i that belong to the cohort c observed in the sample at each time t are added.

Thus, we have the defined subgroups $c = 1,\ldots,C.$ in which n_c is the number of observations within the subgroup c. The size of the cohorts of individuals n_c is important, and it depends directly on the number of cohorts established, C. A larger number of observations in each cohort guarantees greater consistency in the estimation but, in turn, it implies greater heterogeneity within the subgroup. Therefore, the choice of n_c generates a trade-off between homogeneity and robustness.

The set of individuals at time t is defined as $N = C * n_c$, whereas the data set in repeated sections is $S = C * n_c * T * P = N * T * P$.

In general terms, what we are trying to obtain is the expectation of the variable of interest for the cohort c at time t. This is $y^*_{ct} = E\left(y_{it} \mid i \in c,t\right)$. Each variable will also be a conditional expectation of the cohort at each moment, that is, for each variable p we have $z^*_{ct} = E\left(z_{it} \mid i \in c,t\right)$.

If we use the standard panel model presented in (2), the pseudo-panel model results in the following relationship [15]:

$$\begin{aligned} y^*_{ct} = x^*_{ct}\beta + \alpha^*_c + \varepsilon^*_{ct} \\ c = 1,\ldots,C. \\ t = 1,\ldots,T. \end{aligned}$$

However, in practice, the true values of both y^*_{ct} and x^*_{ct} for population cohorts are unobservable, so they are estimated using the sample cohorts observed through $\bar{y}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{it}$ and $\bar{x}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c} x_{it}$.

In this sense, the pseudo-panel model for means of observed cohort samples is expressed as

$$\bar{y}_{ct} = \bar{x}_{ct}\bar{\beta} + \bar{\alpha}_c + \bar{\varepsilon}_{ct}, \tag{3}$$

where $\bar{y}_{ct}$ is the variable of interest mean for the cohort c at time t, and $\bar{\alpha}_c$ is the fixed effects mean at the cohort level for those sample members.

The vector β is obtained from centering each cohort with respect to the average value observed $(\bar{y}_{ct} - \bar{y}_c)$ and the subsequent application of the ordinary least squares method (OLS). According to Guillerm [14], we have, as a result, the following fixed effects or *intra-groups* estimator:

$$\hat{\beta}_{pp} = \left[\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)'(\bar{x}_{ct} - \bar{x}_c)\right]^{-1}\sum_{c=1}^{C}\sum_{t=1}^{T}(\bar{x}_{ct} - \bar{x}_c)'(\bar{y}_{ct} - \bar{y}_c).$$

The fixed effect estimator that is not observed for the mean of the cohort population, α_c, deduced from obtaining $\hat{\beta}_{pp}$, results in

$$\hat{\alpha}_c = \bar{y}_c - \bar{x}_c\hat{\beta}_{pp}.$$

The pseudo-panel approach allows overcoming or, at least, attenuating many of the difficulties that arise in panel models. By using this approach, we can monitor the cohorts over time in repeated cross sections, generating time series for the subgroup means, which can be used as if the panel data were available. This greatly attenuates the bias derived from measurement errors [2], and if the size of the cohorts is large enough $(n_c \to \infty)$ then $\bar{\varepsilon}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c}\varepsilon_{it} \xrightarrow{P} E(\varepsilon_{it}) = 0$ [21], and this ensures consistent parameters [18].

Good results are achieved with the pseudo-panel models [7]. However, they have some limitations, such as the fact that the subsample of individuals to estimate the true values of y^*_{ct} and x^*_{ct} may not be representative [15]. In addition, it is necessary that the sample sizes of the cohort be large enough so that the average of the sample fixed effects $\bar{\alpha}_c$ be a good approximation of the unobserved population mean fixed effect of the cohort α^*_{ct} [7]. Since the individuals observed at each moment are not the same, the average of fixed effects $\bar{\alpha}_c$ can vary when, theoretically, it must be constant [15]. Although the pseudo-panel model can reduce endogeneity, it remains a problem [18].

3. CLASSIFICATION METHOD

Cluster analysis or typical clustering is a technique whose aim is to group or classify a set of data into groups (clusters) of similar objects. Consequently, the clusters are data clusters whose criteria require that the elements within each group be more homogeneous regarding the analyzed characteristics, in relation to the objects of other groups [8]. To determine the similarity among these elements according to the variables involved, mathematical metrics that do not take into account the temporal aspect are used.

Cluster-longitudinal techniques combine content similarities and temporal adjacency in a single representation. This implies using algorithms that consider the temporal "neighbors" of the objects to extract useful knowledge and to be able to analyze the dynamics and structure of the problem under study as completely as possible, as well as the relationship between their multiple factors and determinants.

C. Genolini and B. Falissard [12] and C. Genolini, B. Falissard and J-B. Pingault [13] developed the non-parametric k-means algorithm that allows working with simple or joint trajectories[2].

S is a set of subjects C (cohorts of individuals or pseudo-panels) on which attributes or variables P are measured at different times T. That is,

$$S = C * T * P.$$

Each cohort of individuals is represented by c in which $c = 1,\ldots,C.$, and C is the total number of cohorts, $C = \frac{N}{n_c}$; the subscript $t = 1,\ldots,T.$ indicates each of the different times where the measurements of the variables P of interest $y_1,\ldots,y_P$ are made on the observations, and the subscript p of y_p indicates the attribute being measured, $p = 1,\ldots,P.$.

The sequence made up by the measurements of p on a cohort c at different times T, $y_{c.p} = (y_{c1p}, y_{c2p}, \ldots, y_{cTp})$, is called *simple trajectory* of the variable p in the cohort c.

The first measurement subscript in y_{ctp} refers to the cohort; the second, to time t; and the last one indicates the characteristic or attribute studied.

Consequently, the succession of p in the pseudo-panels C at different times T, $Y_{..p}$, constitutes the *simple trajectory* of the characteristic p for the set of observations $C = \frac{N}{n_c}$.

The series $Y_{..1}, Y_{..2}, \ldots, Y_{..P}$ of the simple trajectory variable P calculated at different times T on the subgroups C is called *joint trajectory* for the set of observations C.

Therefore, we can define each subject c as the matrix $Y_{c..}$ with dimension $P \times T$ of its *joint trajectory* $Y_{c.1}, Y_{c.2}, \ldots, Y_{c.P}$:

$$Y_{c..} = \begin{pmatrix} y_{c11} & y_{c21} & \cdots & y_{cT1} \\ y_{c12} & y_{c22} & \cdots & y_{cT2} \\ \vdots & \vdots & \ddots & \vdots \\ y_{c1P} & y_{c2P} & \cdots & y_{cTP} \end{pmatrix} \qquad (4)$$
$$c = 1,\ldots,C.$$
$$t = 1,\ldots,T.$$
$$p = 1,\ldots,P.$$

The rows of the matrix (4), $Y_{c.p} = \begin{pmatrix} y_{c1p} & y_{c2p} & \ldots & y_{cTp} \end{pmatrix}$, indicate each of the *simple trajectories* P in the cohort c. That is, each row P reflects the temporal evolution of the attribute or characteristic analyzed for each subject.

The columns of the matrix (4), $Y_{ct.} = \begin{pmatrix} y_{ct1} \\ y_{ct2} \\ \vdots \\ y_{ctp} \end{pmatrix}$, indicate the *state of the cohort c* at each different time t. That is, each column T indicates the situation or state of the attributes of each subject at a given time, such as in cross-sectional analyses.

The dataset in repeated sections (S) consists of matrices C $Y_{c..}$. This is: $S = C * Y_{c..}$.

Our goal is to divide the set S into sub-groups M of cohorts of individuals. The condition that the grouping must fulfill is that, according to the characteristics temporarily analyzed, each of the sub-groups m (with $m = 1,\ldots,M$) be created in such a way that the cohorts that are part of it results more homogeneous with respect to the cohorts that are part of the other subgroups.

[2]The algorithm is implemented in the R language with the "k-means for longitudinal data" package developed by the same authors, which was used in this article with the functions "klm" for simple trajectories and "kml3" for joint trajectories.

To measure the similarity or distance d between two cohorts of individuals $Y_{1..}$ and $Y_{2..}$ a metric or distance function Dist is defined and the Euclidean norm $\|.\|$ is used.

According to Genolini et al. [10], the "k-means for longitudinal data" (kml) approach considers two methods to measure this distance d. The first method calculates the distance d^t between two observations $Y_{1..}$ and $Y_{2..}$ taking into account the *state of the cohorts* at different times t. The second method calculates the distance d^p between two observations $Y_{1..}$ and $Y_{2..}$ considering the *simple trajectory* for each characteristic p.

Thus, according to the first method, for each fixed time t we define the distance d^t between $Y_{1..}$ and $Y_{2..}$ as $d_{t.}(Y_{1t.}, Y_{2t.}) = \text{Dist}(Y_{1t.}, Y_{2t.})$. This is the distance between the column t of the matrix $Y_{1..}$ and the column t of the matrix $Y_{2..}$, that is, the distance between the *states of the cohort* at different times t.

The result is a vector of distances T between the two cohorts of individuals:

$$(d_{1.}(Y_{11.}, Y_{21.}), d_{2.}(Y_{12.}, Y_{22.}), \ldots, d_{T.}(Y_{1T.}, Y_{2T.})).$$

Combining the distances T by the function of the norm $\|.\|$ we get the distance between $Y_{1..}$ and $Y_{2..}$ using this first method:

$$d^t(Y_{1..}, Y_{2..}) = \|(d_{1.}(Y_{11.}, Y_{21.}), d_{2.}(Y_{12.}, Y_{22.}), \ldots, d_{T.}(Y_{1T.}, Y_{2T.}))\|.$$

Using the second method for each fixed variable p we define the distance d^p between $Y_{1..}$ and $Y_{2..}$ as $d_{.p}(Y_{1.p}, Y_{2.p}) = \text{Dist}(Y_{1.p}, Y_{2.p})$. This is the distance between the row p of the matrix $Y_{1..}$ and the row p of the matrix $Y_{2..}$, which is the distance between the *simple trajectory* or the temporal evolution of the variable p between the subjects.

The result is a vector of distances P between the two cohorts of individuals:

$$(d_{.1}(Y_{1.1}, Y_{2.1}), d_{.2}(Y_{1.2}, Y_{2.2}), \ldots, d_{.p}(Y_{1.p}, Y_{2.p})).$$

Combining the distances P by the function of the norm $\|.\|$ we get the distance between $Y_{1..}$ and $Y_{2..}$ using the second method:

$$d^p(Y_{1..}, Y_{2..}) = \|(d_{.1}(Y_{1.1}, Y_{2.1}), d_{.2}(Y_{1.2}, Y_{2.2}), \ldots, d_{.p}(Y_{1.p}, Y_{2.p}))\|.$$

According to Genolini et al. [10], if the distance chosen is that of Minkowski for both the first method $\sqrt[a]{\sum_{t=1}^{T} |Y_{1tp} - Y_{2tp}|^a}$ and the second method $\sqrt[a]{\sum_{p=1}^{P} |Y_{1tp} - Y_{2tp}|^a}$, then the distances are equivalent: $d^t(Y_{1..}, Y_{2..}) = d^p(Y_{1..}, Y_{2..})$.

The "kml" package contains a series of criteria for determining the correct number of groups. However, we focus on the criteria of T. Caliński and J. Harabasz [4].

If c_m is the number of cohorts in the group m, then the partition of the elements C can be expressed as $C = c_m * M$.

If $\bar{y}_m$ denotes the average trajectory of the variable y of the cluster m; $\bar{y}$ is the average trajectory of the variable y in the set $S = (c_m * M) * P * T$; and y_{mc} represents the trajectory of y for the cohort c in the m group, then the matrix of sum of squares between groups is defined as

$$B = \sum_{m=1}^{M} c_m (\bar{y}_m - \bar{y})(\bar{y}_m - \bar{y})'.$$

The matrix of sum of squares within groups is

$$W = \sum_{m=1}^{M} \sum_{k=1}^{n_m} (y_{mk} - \bar{y}_{mk})(y_{mk} - \bar{y}_{mk})'.$$

The criteria for measuring homogeneity, which are generally used in the k-means, suggest the minimization of the matrix W. That is, by minimizing the variability of the trajectories, more homogeneous groups are obtained. Using the same reasoning, maximization of matrix B would result in more differentiated groups.

The criteria of Calinski and Harabasz (CH) determine the value M as the number of groups that maximizes the following relationship:

$$CH(M) = \frac{\text{trace}(B)}{\text{trace}(W)} \frac{c-M}{M-1}.$$

4. APPLICATION

4.1. **Materials.** The methodologies developed will be based on the data related to the national Permanent Household Survey (PHS) program[3]. The time period analyzed, from 1989 to 1995 included, is consistent with a time when macroeconomic and structural reforms were promoted in the country and they had a direct impact on the conditions of the labor market. The study is carried out by selecting, from the user databases (WB), salaried workers between 15 and 60 years old from the urban agglomeration Gran Córdoba at the beginning of the time period (1989). During that time, the PHS had a modality of data collection at a specific time with two annual measurements carried out in May and October. This is equal to 13 cross-sectional household data samples. The first sample is from 1989, which is taken as a reference sample. Then, the series is completed with two samples in each of the following six years. Free RStudio software was used to process and analyze these samples[4].

4.2. **Pseudo-panel model.** Data at the level of individuals were added for the building of pseudo-panels, that is, cohorts of individuals who have similar characteristics called *variables factors*. The characteristics selected in this article were the year of birth (in simple ages) and the gender of individuals.

The studied population is salaried workers from the urban agglomeration, for which 92 cohorts of individuals (46 in each gender) were chosen at each time period.

At each time t and in each cohort c, mean variables $\{\bar{y}_{ct1}, \bar{y}_{ct2}\}$ were defined for the subsequent trajectories variables $\{y_{c,.,1}, y_{c,.,2}\}$; the last ones are the pseudo-panel model predictions.

The variable $\bar{y}_{ct1}$ indicates the *mean informality rate* of each cohort c at time t[5]. The mean for each cohort c at time t results in $\bar{y}_{ct1} = \frac{1}{n_c} \sum_{i=1}^{n_c} y_{it}$.

The variable $\bar{y}_{ct2}$ indicates the mean of the income received per hour in the main occupation in each cohort c standardized in the range zero to one. Standardization was performed to avoid the scale problem among the variables involved. That is, each cohort c at time t results in $\bar{y}_{ct2} = \frac{1}{n_c} \sum_{i=1}^{n_c} y^*_{it2}$, where y_{it2} is the average income per hour of worker i at time t and $y^*_{it2} = \frac{(y_{it2} - mn(y_{it2}))}{(Mx(y_{it2}) - mn(y_{it2}))}$ is the standardized average income per hour of worker i at time t.

[3]This program is jointly carried out by the National Statistics and Censuses Institute of Argentina (Spanish: INDEC) and the Provincial Statistics Bureau (Spanish: DPE).

[4]The longitudinal clustering technique was chosen using the "kml3d" package. Repository CRAN R-project. Collate global.r distance 3d.rclusterLongData3d.r kml3d.r Available from `https://cran.r-project.org/web/packages/kml3d/index.html`

[5]The condition of "labor informality" was defined as the total or partial rejection of any of the following rights and/or benefits in the population of salaried workers: dismissal compensation, paid time off, mid-year and end-year bonus, work insurance, retirement discount.

The trajectory $y_{c,.,1}$ was obtained from $\hat{\bar{y}}_{ct1}$, which was based on a set of regressor variables $\bar{x}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c} x_{it}$. The set $(\bar{x}_{ct})$ is made up of the means of the explanatory variables complete secondary education level (x_1), working hours (x_2), seniority in the job position (x_3), company size (x_4) and age (x_5) in each cohort c at time t.

The trajectory $y_{c,.,2}$ was obtained from $\hat{\bar{y}}_{ct2}$, which was based on its standardized mean value $(\bar{y}_{ct2})$. That is, $\hat{\bar{y}}_{ct2} = \bar{\beta} + \bar{\alpha}_c + \bar{\varepsilon}_{ct}$. Table 1 shows the results obtained.

TABLE 1. Results of the pseudo-panel model.

Informality	**Coef.**	**Std. Err.**	**t**	$P > \|t\|$
Secondary	−0.1207207	0.0517498	−2.33	0.020
Hours	0.1473925	0.0427889	3.44	0.001
Seniority	−0.2263262	0.0450433	−5.02	0.000
Company size	0.4217886	0.0395092	10.68	0.000
Age	−0.0104426	0.0027331	−3.82	0.000
Constant	0.8077621	0.1131973	7.14	0.000
σ_u	0.1310602	F test $(H_0 : u_i)$ = 2,97		
σ_e	0.17214083	Prob $> F$ = 0.0000		
ρ	0.36695266	Groups = 92		N = 11.760

4.3. **Temporal classification model.** Once the income trajectories $y_{c,.,1}$ and the informality rate $y_{c,.,2}$ have been estimated, each of the 92 cohorts is expressed as the following matrix of their joint trajectories:

$$Y_{c..} = \begin{pmatrix} y_{c,1,1} & y_{c,2,1} & \cdots & y_{c,13,1} \\ y_{c,1,2} & y_{c,2,2} & \cdots & y_{c,13,2} \end{pmatrix}.$$

While the metric or distance function Dist was the Euclidean one, the criterion for the optimization of groups was that of Calinski and Harabasz.

Figure 1 shows the results of the correct number of clusters and the simple trajectories of both variables. In these results there is a superposition between the mean trajectories of groups (clusters) made up on simple trajectories of income $y_{c,.,1}$ and the informality rate $y_{c,.,2}$ of the cohorts. While Table 2 indicates the cohorts that make up each of the groups, Figure 2 shows the joint trajectories of the "centers" of trajectories of clusters.

Figure 2 shows that the clusters present important differences in levels of informality rate and income per hour, and present similar tendency joint trajectories. However, the tendency slopes are a little different between the clusters, and consequently their trajectories do not intersect. This indicates that the differences between groups are roughly the same for each of the time periods.

4.4. **Results.** The exploratory analysis allows identifying relationships among the variables and the analyzed cohorts, as well as drawing potential conclusions in relation to the phenomenon under study. According to the joint trajectories studied, the correct number of clusters was three (A, B, C). Group A (blue) is characterized by high levels of informality and relatively low mean labor income rates per hour compared to other groups. This

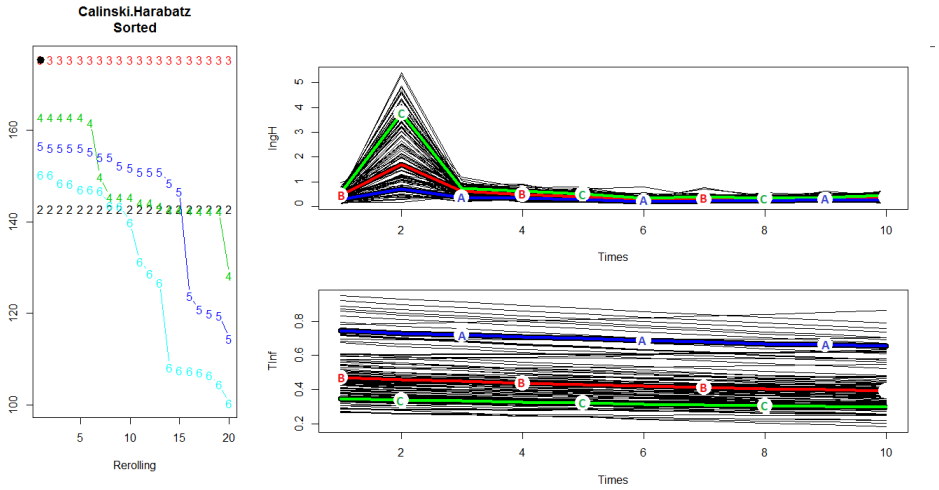

FIGURE 1. Longitudinal clustering. Calinski and Harabasz criterion and simple trajectories of income and informality.

TABLE 2. Results of longitudinal clustering.

Cluster	**Cohorts**		n *(%)*
A	i_v15 to i_v19	i_m15 to i_m23 i_m53 to i_m60	22 *(23,9)*
B	i_v20 to i_v27	i_m24to i_m52	37 *(40,2)*
C	i_v28 to i_v60		33 *(35,9)*

group A can be defined as the one with the greatest socio-economic "vulnerability". Within it, two demographic sub-groups are distinguished. The first one is integrated by the cohorts that, at the beginning of the time period, entered the labor market: men between 15 and 19 years old and women between 15 and 23 years old. The second subgroup is made up of the cohorts of women who were between 53 and 60 years old (in 1989). Group B (red) is characterized by a greater female composition. This includes cohorts which, in 1989, included people between 20 and 27 years old in the case of men and between 24 and 52 years old in the case of women. That is, men in the first years of "stability" and women in years of "employability" in terms of labor market. Taking into account the first group, the mean rates of informality are relatively lower, whereas the mean trajectory of the labor income per hour is higher, even in a substantially smaller proportion. Group C (green) is only characterized by male members. Cohorts that are part of this group included people between 28 and 60 years old at the beginning of the time period (1989). That is, men in ages of greater "employability" and in ages towards the end of the labor market. This group is characterized by low informality rates and income levels much higher than the other two defined groups. In this sense, it could be defined as the one with better socio-economic conditions. Regarding the characteristics used in the classification, a potential relationship between informality levels and earned labor income could be considered. The levels which

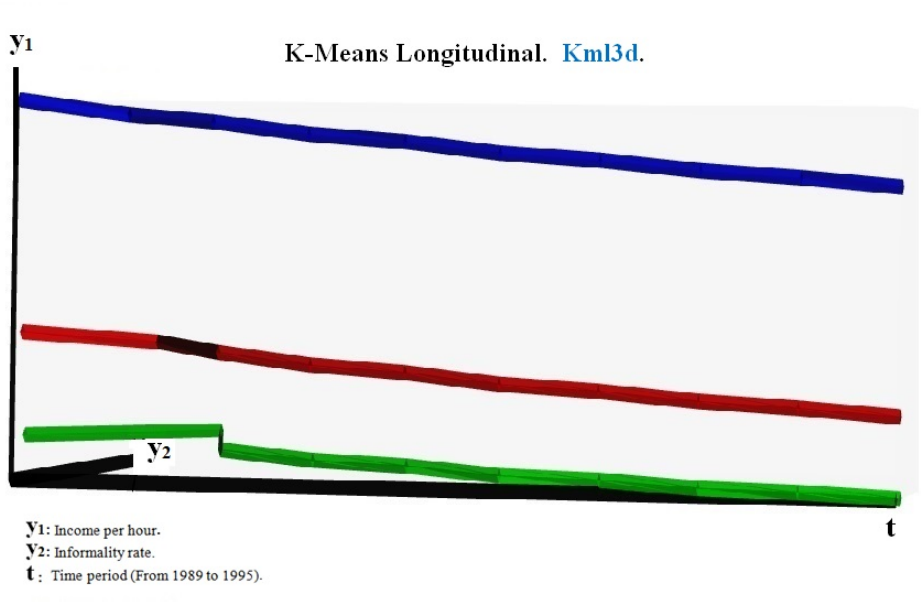

FIGURE 2. Longitudinal clustering. Mean trajectories of the groups.

have these variables could also be related to the gender of the cohort. Indirectly, the results may give rise to some hypotheses: it takes more time for women than for men to enter the formal labor market (Group A) and these cohorts do not integrate the groups that receive a higher mean income (Group C).

5. CONCLUSIONS

In Argentina, informal economy did not represent significant levels until the end of the first half of 1970. The economic reforms that have been carried out in Latin American countries since the mid-1980s and their impact on labor markets caused a decrease in the quality of employment [22]. Therefore, during the 1990s, informal employment in urban areas of Latin America grew from approximately 50% to 58% [9]. Consequently, the study of the evolution of informal employment profiles over time becomes relevant due to the conditions of "vulnerability" to which these workers are exposed.

The pseudo-panel approach allows overcoming the lack of availability of temporal data by the creation of variables factors that define cohorts of individuals who have characteristics in common. In this article, we focused on year of birth and gender of individuals.

Some studies in Argentina have used pseudo-panel for studying poverty [6, 19], income mobility [20], and particularly labor informality [1], where this methodology was used to compare the informality and unemployment series evolution.

In this article, the pseudo-panel model provides the necessary information to define profiles of informal workers that combine similarities in the evolution of their characteristics through a longitudinal *k*-means approach.

The combination of these methodologies is a contribution to the study in the area of social and economic development, which allows identifying relationships between the variables and the cohorts analyzed as well as drawing potential conclusions in relation to the phenomenon under study when there are no complete panels.

With the exploratory analysis applied in this article we could identify three profiles of well-defined informal workers taking into account the temporal trajectories of their informality and income levels.

REFERENCES

[1] O. Arias and W. S. Escudero, *Assessing trends in informality in Argentina: A cohorts panel VAR approach*, Mimeografía, Banco Mundial y CEDLAS, 2007.

[2] F. Antman and D. McKenzie, *Earnings mobility and measurement error: A pseudo-panel approach*, Economic Development and Culture Change **56** (2007), no. 1, 125–161. https://doi.org/10.1086/520561.

[3] M. Arellano y O. Bover, *La econometría de datos de panel*, Investigaciones Económicas (Segunda época), **14** (1990), no. 1, 3–45.

[4] T. Caliński and J. Harabasz, *A dendrite method for cluster analysis*, Communications in Statistics **3** (1974), no. 1, 1–27. https://doi.org/10.1080/03610927408827101.

[5] G. Canavire-Bacarreza, J. A. Urrego and F. Saavedra, *Informality and mobility in the labor market: A pseudo-panel's approach*, Revista Latinoamericana de Desarrollo Económico, no. 27 (2017), 57–75. Available from http://www.scielo.org.bo.

[6] L. Casanova, *Trampas de Pobreza en Argentina: Evidencia Empírica a Partir de un Pseudo Panel*, Documento de Trabajo no. 64, La Plata: CEDLAS, 2008. Available from http://sedici.unlp.edu.ar/.

[7] A. Deaton, *Panel data from time series of cross-sections*, Journal of Econometrics **30** (1985), no. 1-2, 109–126. https://doi.org/10.1016/0304-4076(85)90134-4.

[8] M. Garre, J. J. Cuadrado, M. A. Sicilia, D. Rodriguez y R. Rejas, *Comparación de diferentes algoritmos de clustering en la estimación de coste en el desarrollo de software*, Revista Española de Innovación, Calidad e Ingeniería del Software **3** (2007), no. 1, 6–22. https://www.redalyc.org/articulo.oa?id=92230103

[9] L. Gasparini y L. Tornarolli, *Labor Informality in Latin American and the Caribbean: Patterns and Trends from Household Surveys Microdata*, Documento de Trabajo no. 46, La Plata: CEDLAS, 2007. Available from http://sedici.unlp.edu.ar/.

[10] C. Genolini, X. Alacoque, M. Sentenac and C. Arnaud, *kml and kml3d: R packages to cluster longitudinal data*, Journal of Statistical Software **65** (2015), no. 4, 1–34. https://doi.org/10.18637/jss.v065.i04.

[11] C. Genolini and B. Falissard, *KmL: k-means for longitudinal data*. Computational Statistics **25** (2010), no. 2, 317–328.

[12] C. Genolini and B. Falissard, *Package 'kml'*, 2016. https://cran.r-project.org/web/packages/kml/kml.pdf.

[13] C. Genolini, B. Falissard and J.-B. Pingault, *Package 'kml3d'*, 2017. https://cran.r-project.org/web/packages/kml3d/kml3d.pdf.

[14] M. Guillerm, *Les méthodes de pseudo-panel*, Document de travail. Institut National de la Statistique et des Etudes Economiques. París, 2015. Available from https://www.insee.fr.

[15] M. Guillerm, *Pseudo-panel methods and an example of application to Household Wealth data*, Economie et Statistique, no. 491-492 (2017), 109–130. https://doi.org/10.24187/ecostat.2017.491d.1908.

[16] J. A. Hausman, *Specification tests in econometrics*, Econometrica **46** (1978), no. 6, 1251–1271. https://doi.org/10.2307/1913827.

[17] ILO, *Mujeres y hombres en la economía informal: un panorama estadístico (tercera edición)*, Ginebra: OIT, 2018. Available from https://www.ilo.org.

[18] Y. Meng, A. Brennan, R. Purshouse, D. Hill-McManus, C. Angus, J. Holmes and P. S. Meier, *Estimation of own and cross price elasticities of alcohol demand in the UK — A pseudo-panel approach using the Living Costs and Food Survey 2001–2009*, Journal of Health Economics **34** (2014), 96–103. https://doi.org/10.1016/j.jhealeco.2013.12.006.

[19] M. E. Millón and B. R. García, *Trampas de Pobreza: evidencia para las regiones de Argentina*, Atlantic Review of Economics (ARoEc) **1** (2018), no. 2, 40 pp.

[20] A. I. Navarro, *Estimating Income Mobility in Argentina with Pseudo-Panel Data*, Preliminary version, Departamento de Economía, Universidad de San Andrés y Universidad Austral, 2006.

[21] J. M. Perera, *La movilidad de las rentas laborales en el mercado de trabajo uruguayo: Un enfoque de pseudo-panel*, Documentos de Trabajo. CINVE, Montevideo. Junio, 2006. Available from https://cinve.org.uy.

[22] J. Weller, *Reformas económicas y situación del empleo en América Latina*, en: Entre el trabajo y la política: las reformas de las políticas sociales argentinas en perspectiva comparada. Buenos Aires: Biblos, 2003.

[23] J. M. Wooldridge, *Introductory Econometrics: A Modern Approach*, 3rd edition, Mason, OH: Thomson/ South-Western, 2006.

[24] J. M. Wooldridge, *Econometric Analysis of Cross Section and Panel Data*, 2nd edition, Cambridge, Mass.: MIT Press, 2010.

(M. L. Iglesias) UNIVERSITY OF CÓRDOBA, INSTITUTE OF STATISTICS AND DEMOGRAPHY, 5000 CÓRDOBA, ARGENTINA, AND UNIVERSITY OF CÓRDOBA, SECRETARÍA DE CIENCIA Y TECNOLOGÍA, 5000 CÓRDOBA, ARGENTINA

Email address: miglesias@unc.edu.ar

(M. I. Stimolo) UNIVERSITY OF CÓRDOBA, FACULTY OF ECONOMICS SCIENCE, 5000 CÓRDOBA, ARGENTINA, AND CENTRO DE INVESTIGACIONES EN CIENCIAS ECONÓMICAS, CIECS UNC-CONICET, 5000 CÓRDOBA, ARGENTINA

Email address: maria.ines.stimolo@unc.edu.ar