Perfiles de la informalidad laboral en Gran Córdoba

Análisis longitudinal

Iglesias,Maximiliano Luján
Workshop
Instituto de Estadística y Demografía
FCE - UNC

Agosto,2019

Análisis Longitudinal. La *característica* que define a un estudio longitudinal es que los individuos se miden repetidamente a través del tiempo. La *principal ventaja* es su capacidad para separar lo que en el contexto de los estudios de población se llama efectos de *cohorte* y de la *edad*. [1]

[1] Diggle P, Heagerty P, Kung-Yee, L and Scott L (2004)

- Una de las *principales limitaciones*, para el correcto análisis desde un abordaje de la variabilidad temporal del mercado laboral en los países en desarrollo es la escasez de información apropiada de datos de panel disponibles. [2]
- Los *paneles* suelen ser *rotativos* donde los hogares o individuos permanecen un período relativamente corto en la muestra.
- El abandono no aleatorio de ciertas unidades (attrition) puede generar un sesgo considerable en las estimaciones. [3]

[2] Canavire-Bacarreza, Urrego and Saavedra (2017)
[3] Perera,J. (2006)

Objetivo. El presente trabajo estudia perfiles de trabajadores en condiciones de informalidad laboral en Gran Córdoba durante la década del 90' mediante metodologías que permitan la incorporación de la dimensión temporal en el análisis.

Informalidad. Desde la década del 70' se ha venido utilizando el concepto **Economía Informal** de manera generalizada y poco precisa. En los últimos años el concepto ha evolucionado permitiendo una mayor rigurosidad metodológica derivando en dos enfoques. El primero se encuentra centrado en el establecimiento en que se realiza la actividad, dando origen al concepto de **Sector Informal**. El segundo, se enfoca en las condiciones en que se realiza la actividad, dando sustento al concepto de **Empleo Informal**.

En el modelo de panel típico se agrega un efecto fijo individual al modelo lineal estándar, para capturar el efecto de las características individuales que son constantes en el tiempo sobre la variable de interés.[4]

$$y_{it} = X_{it}\beta + \alpha_i + \varepsilon_{it}$$

i="individuos", i=1,...,n., y t="tiempo", t=1,...,T.

[4] Guillerm, M. (2017)

El enfoque de pseudo panel permite superar estas limitaciones mencionadas mediante la construcción de paneles "sintéticos". Esto se logra, reemplazando las observaciones individuales del panel original con medias de subgrupos de la población que se puede identificar su aparición en repetidas encuestas transversales.[5]

$$\bar{y_{c(t)}} = \bar{X_{ct}}\beta + \bar{\alpha_c} + \bar{\varepsilon_{ct}}$$

c="número de grupos", c=1,...,C. y t=1,...,T.

[5] Meng Y, Brennan A, Purshouse R and Hill-McManus D (2014).

- Para definir los subgrupos se utilizan factores que deben ser o suponerse invariantes en el tiempo (por ejemplo, año de nacimiento, género, etnia, etc).

N conjunto de datos en secciones repetidas, **C** número de subgrupos definidos, $\mathbf{n_c}$ número de observaciones dentro del grupo.

$$N = C * n_c * T$$

- Las técnicas de clúster-temporal combinan similitudes de contenido y adyacencia temporal en una sola representación.
- Las k-means for Longitudinal data (Genolini and Falissard) constituyen una implementación de k-means diseñados para funcionar específicamente en trayectorias **(kml)** o en trayectorias conjuntas **(kml3d)**.

- Sea **S** un conjunto de **C** sujetos (cohortes de individuos o pseudo paneles).

$$S = C * T * P$$

- Para cada sujeto **"c"**, tenemos **"p"** variables de resultados $Y_{..A}$,$Y_{..B}$,... ,$Y_{..P}$ medidas en **"t"** momentos.
- La secuencia de la medición de la variable *"A"* $Y_{..A}$ para cada sujeto **"c"** en los **"t"** momentos temporales se denomina "variable trayectoria" de *"A"* para el sujeto **"c"**.

Podemos definir a cada sujeto **"c"** como la matriz P*T de sus *"trayectorias conjuntas"*.

$$Y_{c..} \begin{pmatrix} y_{c1A} & y_{c2A} & \cdots & y_{cTA} \\ y_{c1B} & y_{c2B} & \cdots & y_{cTB} \\ \vdots & \vdots & \cdots & \vdots \\ y_{c1P} & y_{c2P} & \cdots & y_{cTP} \end{pmatrix}$$

$$\begin{aligned} \mathrm{c} &= 1, \ldots, C. \\ \mathrm{t} &= 1, \ldots, T. \\ \mathrm{p} &= A, \ldots, P. \end{aligned}$$

- Cada elemento $\boldsymbol{y_{c..}}$ de la matriz $\boldsymbol{Y_{c..}}$ corresponde al valor de la estimación de la variable "**p**" para la cohorte "**c**" en el momento "**t**".

- Las filas de la matriz $\boldsymbol{Y_{c..}}$ indican la trayectoria de una variable $\boldsymbol{Y_{c.A}} = (y_{c1A} \quad y_{c2A} \quad \cdots \quad y_{cTA})$.

- Las columnas de la matriz $\boldsymbol{Y_{c..}}$ indica el "*estado del individuo*" en el momento "**t**"

$$\boldsymbol{Y_{ct.}} = \begin{pmatrix} y_{ctA} \\ y_{ctB} \\ \vdots \\ y_{ctP} \end{pmatrix}.$$

- El *objetivo* del agrupamiento es dividir **S** en **"m"** sub-grupos homogéneos.
- Para calcular la distancia **"d"** entre el sujeto $Y_{1..}$ y el sujeto $Y_{2..}$
- Kml plantea *dos métodos*: El primer método calcula la distancia **d** entre $Y_{1..}$ y $Y_{2..}$ considerando los *"estados de los individuos"* en cada momento **"t"**. El segundo método calcula la distancia **d'** entre $Y_{1..}$ y $Y_{2..}$ considerando las *"trayectorias"* para cada variable **"p"**.[6]

[6] Genolini C and Falissard B (2010).

- **Método I.**

Para calcular la distancia **d** entre $\boldsymbol{Y_{1..}}$ y $\boldsymbol{Y_{2..}}$ de acuerdo con el *primer método*, para cada **t** fija, definimos la distancia entre $\boldsymbol{Y_{1..}}$ y $\boldsymbol{Y_{2..}}$ (distancia entre el "*estado de los individuos*" en el momento **t**) como $\boldsymbol{d_{t.}(Y_{1t.}, Y_{2t.}) = Dist(Y_{1t.}, Y_{2t.})}$.
Esta es la distancia entre la columna **t** en la matriz $\boldsymbol{Y_{1..}}$ y la columna **t** en la matriz $\boldsymbol{Y_{2..}}$.

El resultado es un "*vector de T distancias*".

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

Luego combinamos esas **t** distancias usando una función que algebraicamente corresponde a una *norma* $\|.\|$ del vector distancia.

Finalmente, la distancia entre $\boldsymbol{Y_{1..}}$ y $\boldsymbol{Y_{2..}}$ es

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

- **Método II.**

Para calcular, según el segundo método, la distancia $\boldsymbol{d'}$ entre $\boldsymbol{Y_{1..}}$ y $\boldsymbol{Y_{2..}}$ para cada variable "**p**", definimos la distancia entre $\boldsymbol{Y_{1.p}}$ y $\boldsymbol{Y_{2.p}}$ (distancia entre dos trayectorias individuales "**p**") como $\boldsymbol{d_{.p}(Y_{1.p}, Y_{2.p}) = Dist(Y_{1.p}, Y_{2.p})}$.

Esta es la distancia entre la línea **p** en la matriz $\boldsymbol{Y_{1..}}$ y la línea **p** en la matriz $\boldsymbol{Y_{2..}}$.

El resultado es un "vector de **p** distancias".

$$\left(d_{.A}(Y_{1.A}, Y_{2.A}), d_{.B}(Y_{1.B}, Y_{2.B}), \ldots, d_{.P}(Y_{1.P}, Y_{2.P})\right)$$

Luego combinamos esas **p** distancias usando una función que algebraicamente corresponde a una *norma* $\|.\|$ del vector distancia.

Finalmente, la distancia entre $\boldsymbol{Y_{1..}}$ y $\boldsymbol{Y_{2..}}$ es

$$\boldsymbol{d'(Y_{1..}, Y_{2..}) = \left\|\left(d_{.A}(Y_{1.A}, Y_{2.A}), d_{.B}(Y_{1.B}, Y_{2.B}), \ldots, d_{.P}(Y_{1.P}, Y_{2.P})\right)\right\|}$$

7

- Para elegir el óptimo de grupos *KmL* utiliza el criterio **"C(g)"** de Calinski y Harabasz.[7]
- Sea $\mathbf{c_m}$ como el número de trayectorias en el grupo **m**.

$$C = c_m * m$$

- $\mathbf{\overline{y}_m}$ la trayectoria media del cluster **"m"**.
- $\mathbf{\overline{y}}$ es la trayectoria media del conjunto de datos **"S"**.

$$S = (c_m * m) * T * P$$

[7] Genolini C and Falissard B (2010).

MATRIZ VARIANZA ENTRE (entre grupos).

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

MATRIZ VARIANZA DENTRO (dentro de grupos).

$$W = \sum_{m=1}^{g} \sum_{k=1}^{n_m} (y_{mk} - \bar{y}_m)\,(y_{mk} - \bar{y}_m)'$$

El **número óptimo** de agrupamientos (clústeres) corresponde al valor **G** que maximiza

$$C(g) = \frac{traza\,(B)}{traza\,(W)} \frac{c-g}{g-1}$$

- Los datos corresponden al programa **Encuesta Permanente de Hogares (EPH - Puntual)** realizado por el Instituto Nacional de Estadísticas y Censos (INDEC) de Argentina entre los años **1989** y **1995** con dos ondas anuales considerándose el aglomerado *Gran Córdoba*.
- Se construyeron *pseudo paneles* para cada uno de los períodos mencionados. Los *factores* utilizados para su construcción fueron el *año de nacimiento*(edades simples) y el *genero* de los individuos.
- Se definió la condición de **"informalidad laboral"** como la denegación total o parcial de alguno de los siguientes derechos y/o beneficios en la población de trabajadores **asalariados**: indemnización por despido, vacaciones, aguinaldo, seguro de trabajo, descuento jubilatorio.

- Se considero como población de análisis a los trabajadores **asalariados** que presentaban edades entre **15** y **60** al inicio del período (1989) determinándose *92 cohortes* de individuados en cada momento del tiempo.
- Se estimo la variable trayectoria de la informalidad bajo dos modelos.
- **Modelo 1**.

$$\bar{y_{ct}} = \frac{1}{n_c} \sum_{i=1}^{n_c} y_{it}$$

- **Modelo 2**.

$$\bar{y_{c(t)}} = \bar{X_{ct}}\beta + \bar{Z_{ct}}b_{ct} + \bar{\varepsilon_{ct}}$$

- **Modelo 3**.

$$\bar{y_{c(t)}} = \bar{X_{ct}}\beta + \bar{Z_{ct}}b_{ct} + \bar{\varepsilon_{ct}}$$

$$\bar{z_{ct}} = \frac{1}{n_c}\sum_{i=1}^{n_c} z_{it}$$

c=1,...,92, t=1,...,10, y="tasa de informalidad" z="ingreso por hora

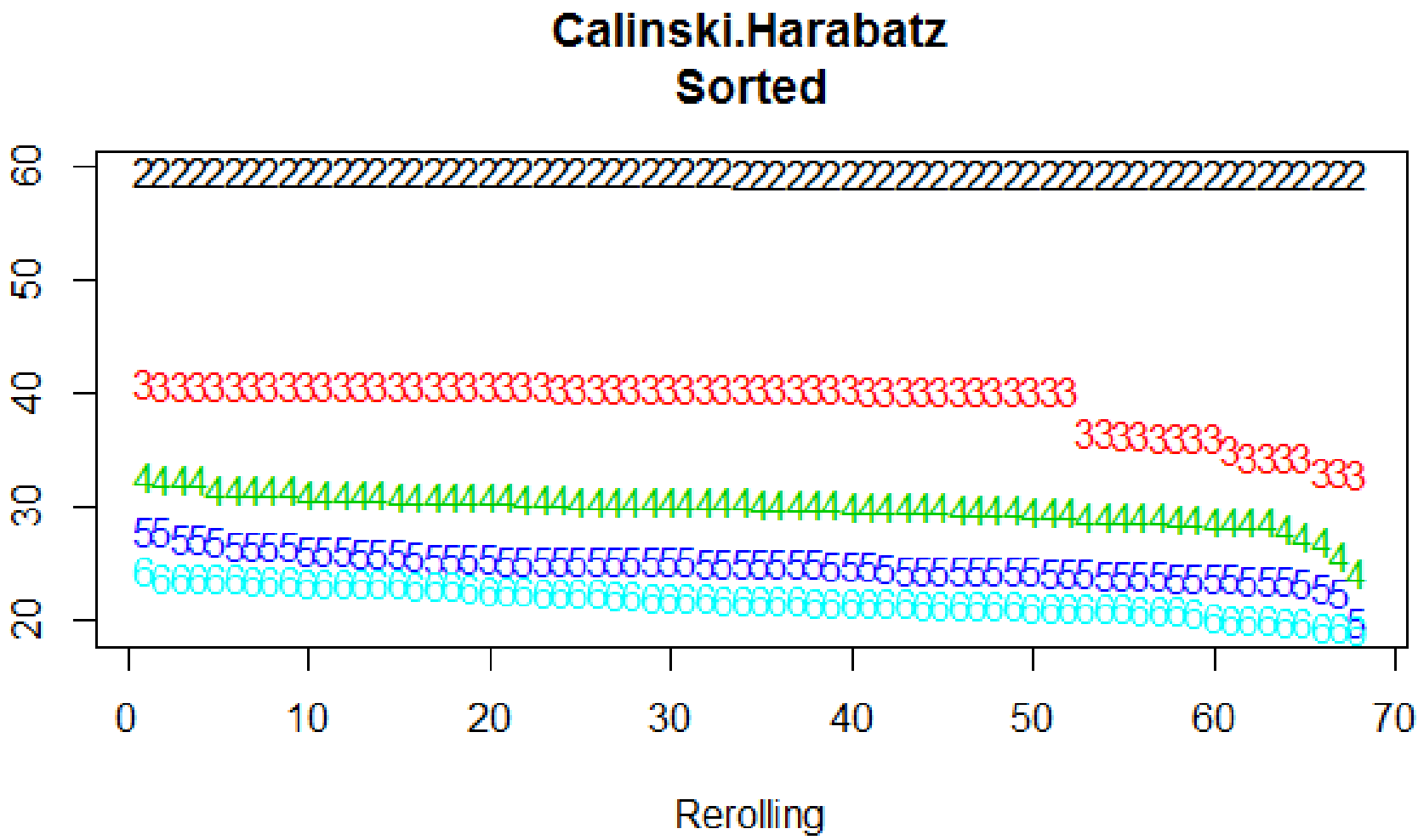
- **Paquetes RStudio**. Para trabajar con trayectorias únicas se utilizo el paquete **'kml'** de *Genolini and Falissard* (Febrero,2016)[8]. Mientras que, para las trayectorias conjuntas se uso **'kml3d'** de *Genolini, Falissard and Pingault*[9] (Agosto,2017).

[8] Repository CRAN R-project. Collate global.R clusterLongData.R parKml.R parChoice.R kml.R

[9] Repository CRAN R-project. Collate global.r distance3d.r clusterLongData3d.r kml3d.r

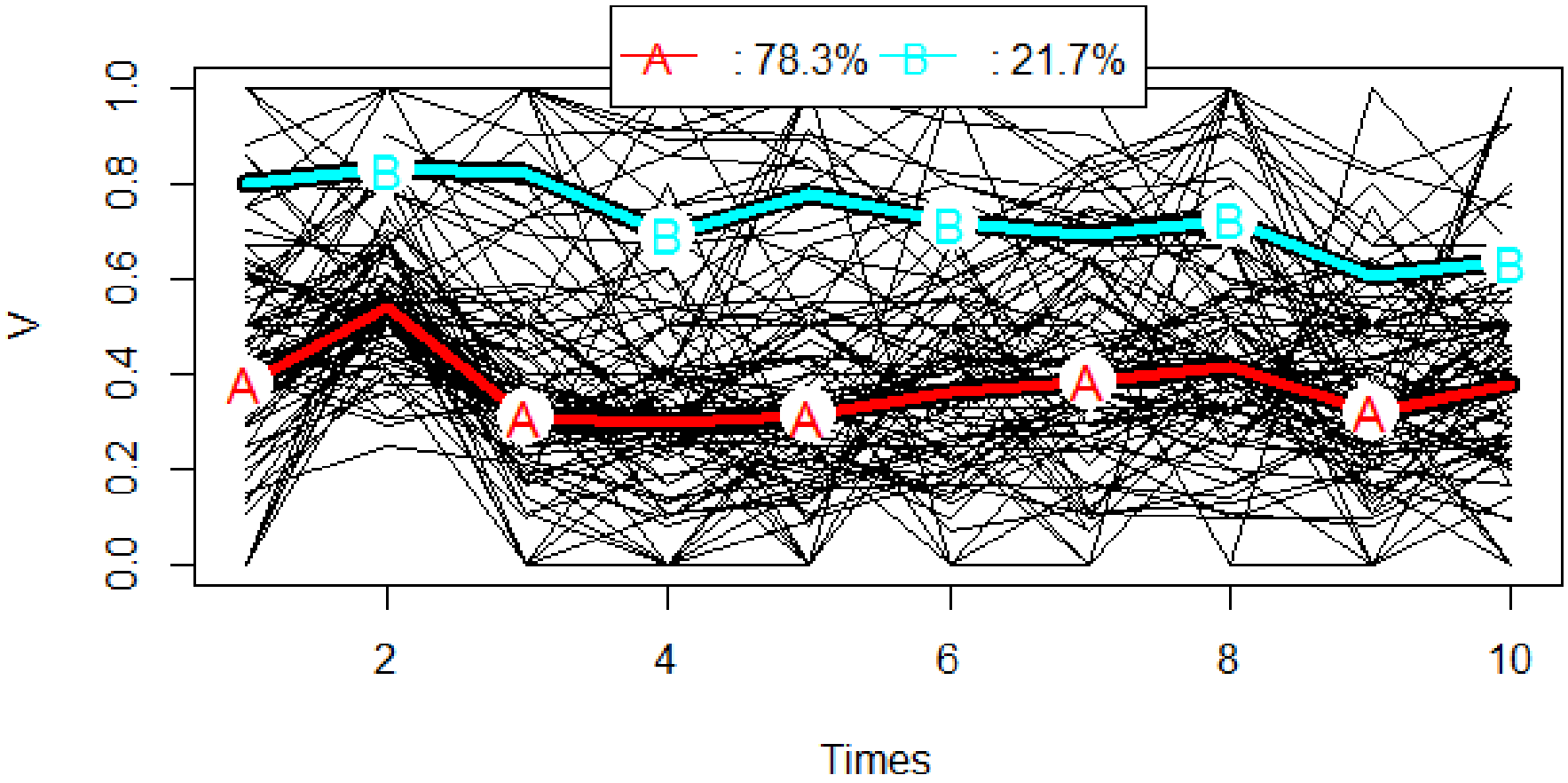
Aplicación

Modelo 1. Caliliski T and Harabasz J



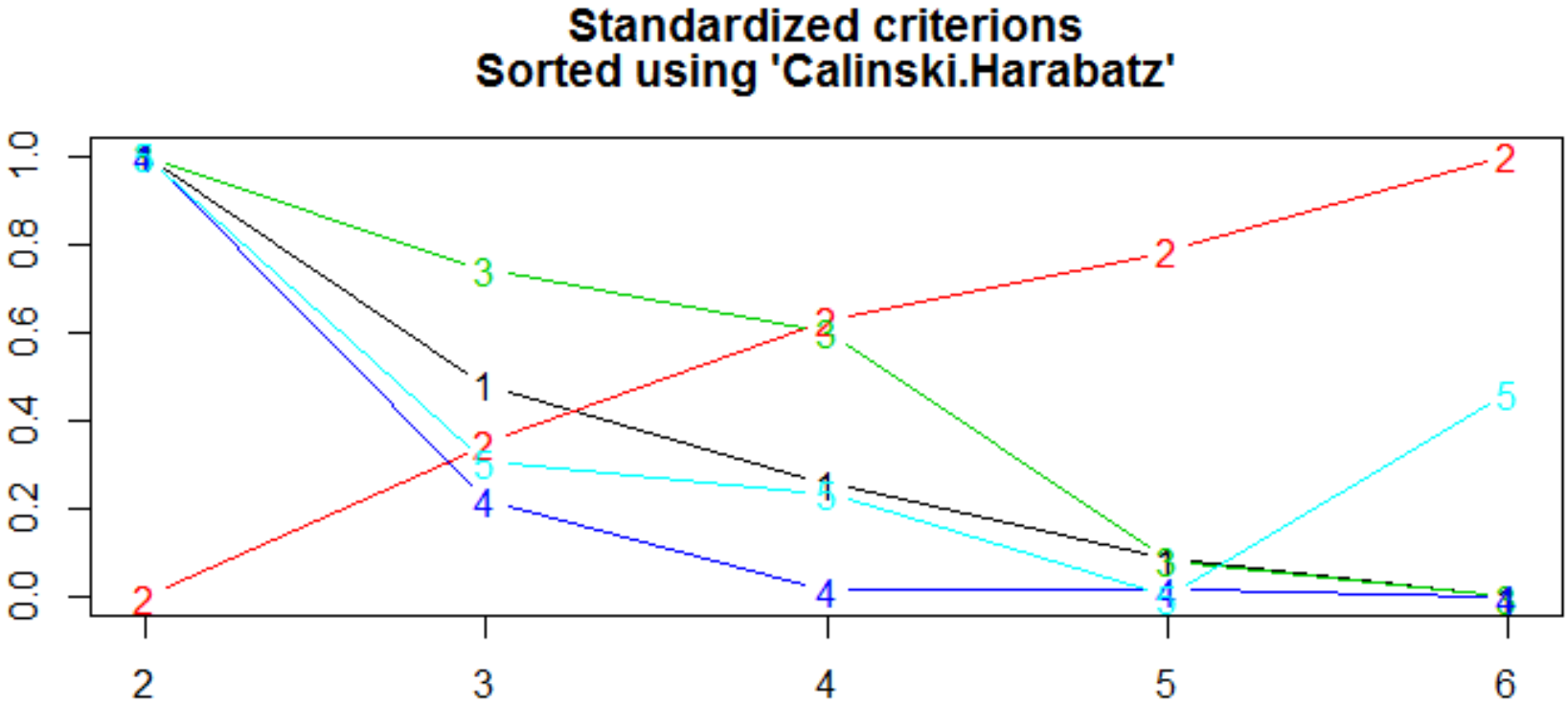
Aplicación

Modelo 1. Trayectorias



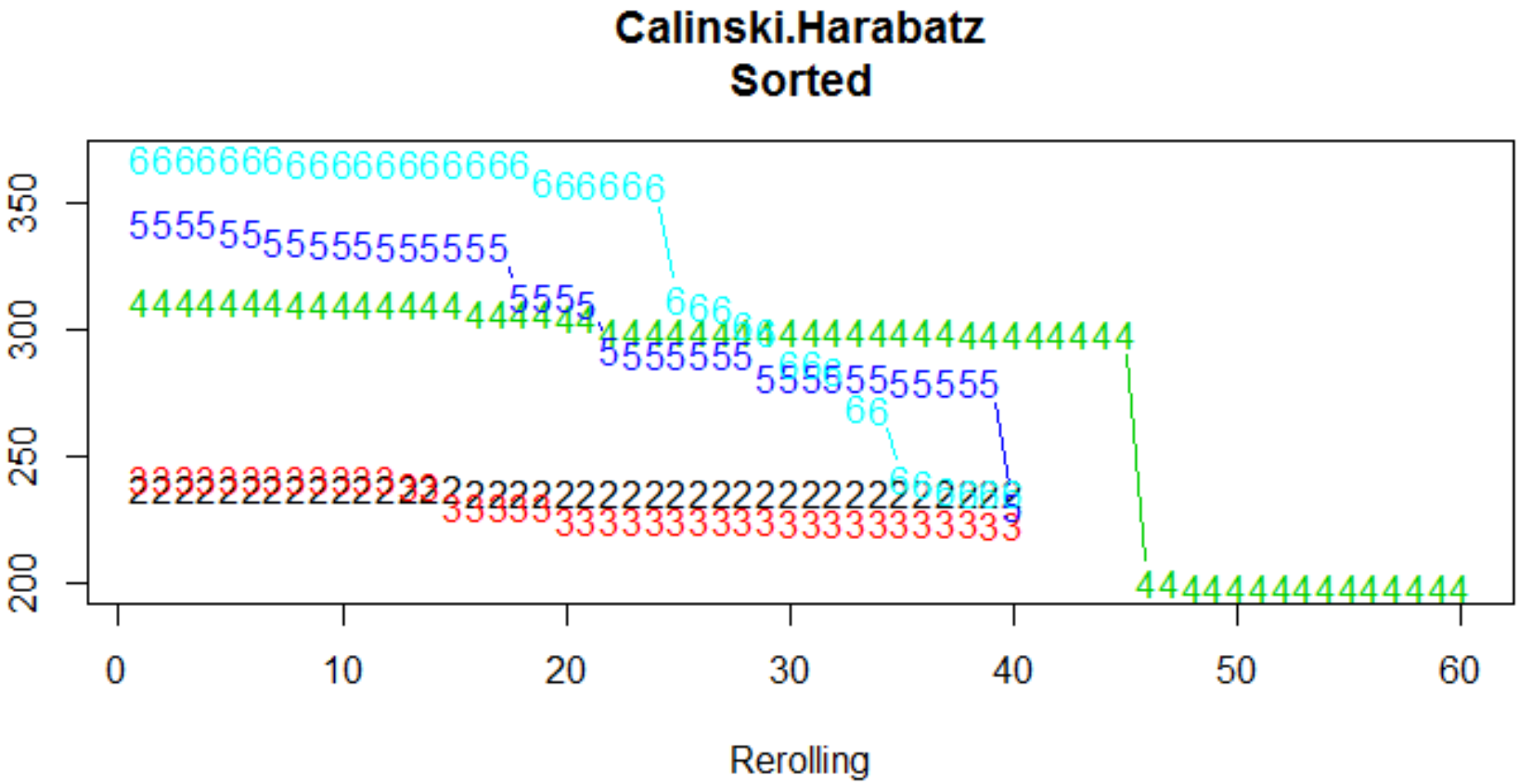
Aplicación

Modelo 1. Criterios de Selección



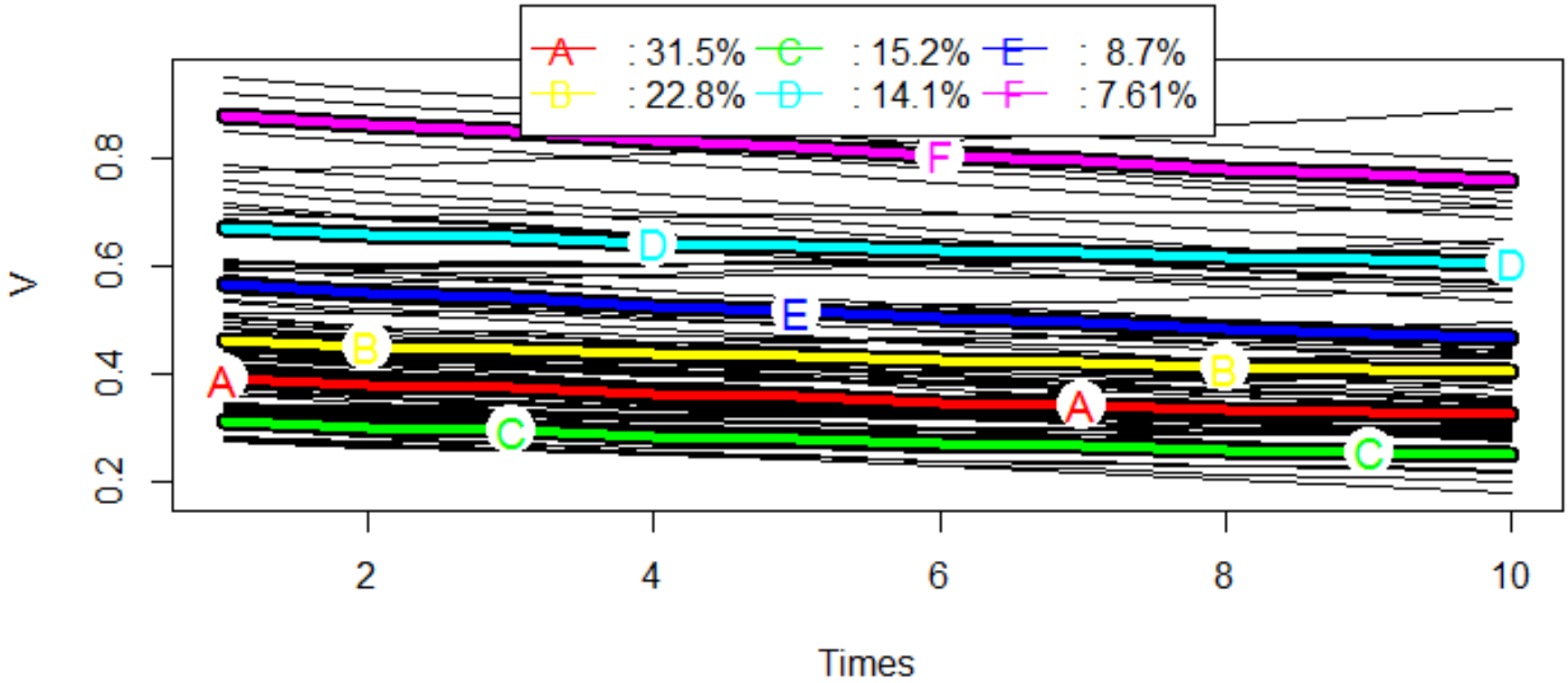
Aplicación

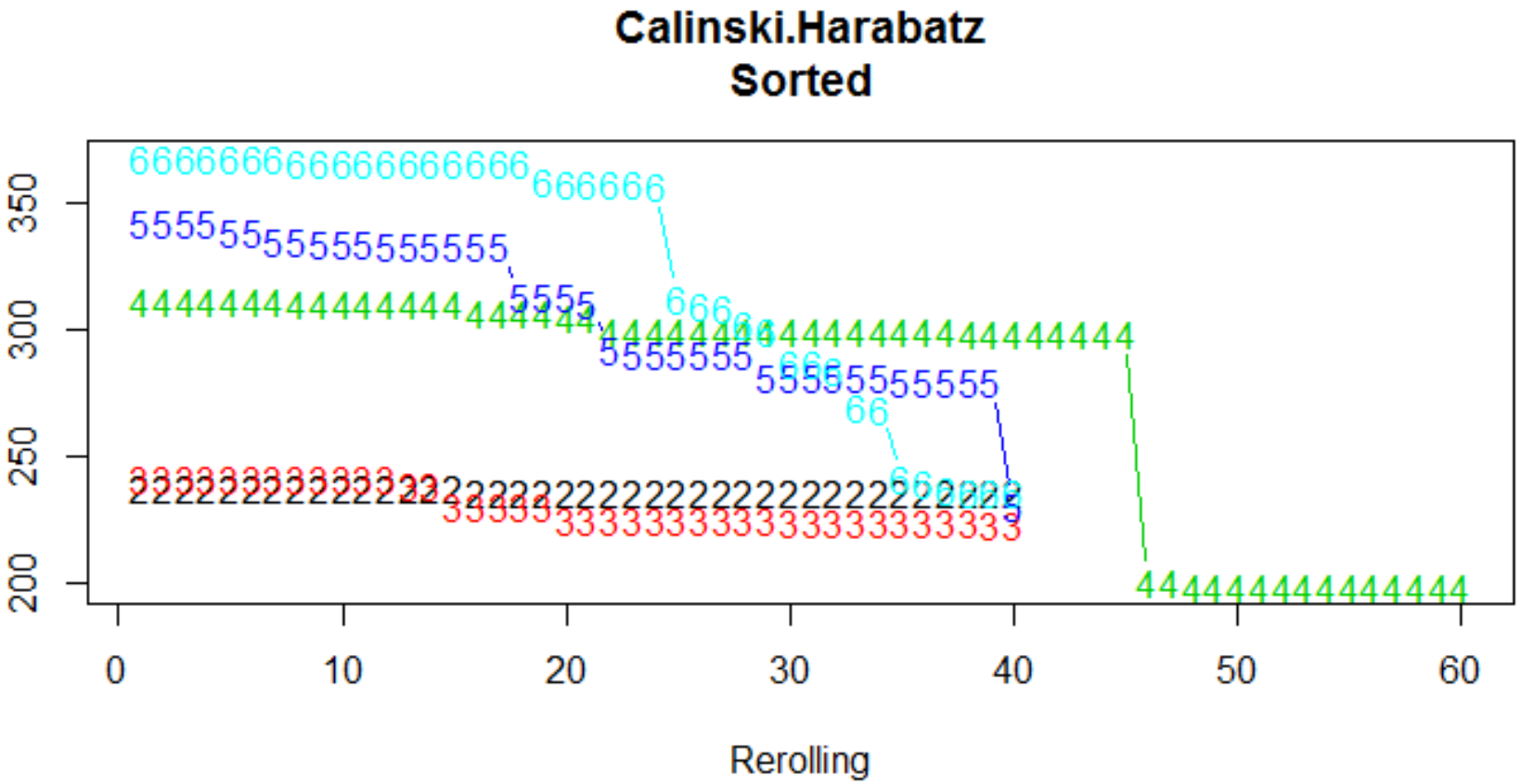
Modelo 2. Caliliski T and Harabasz J

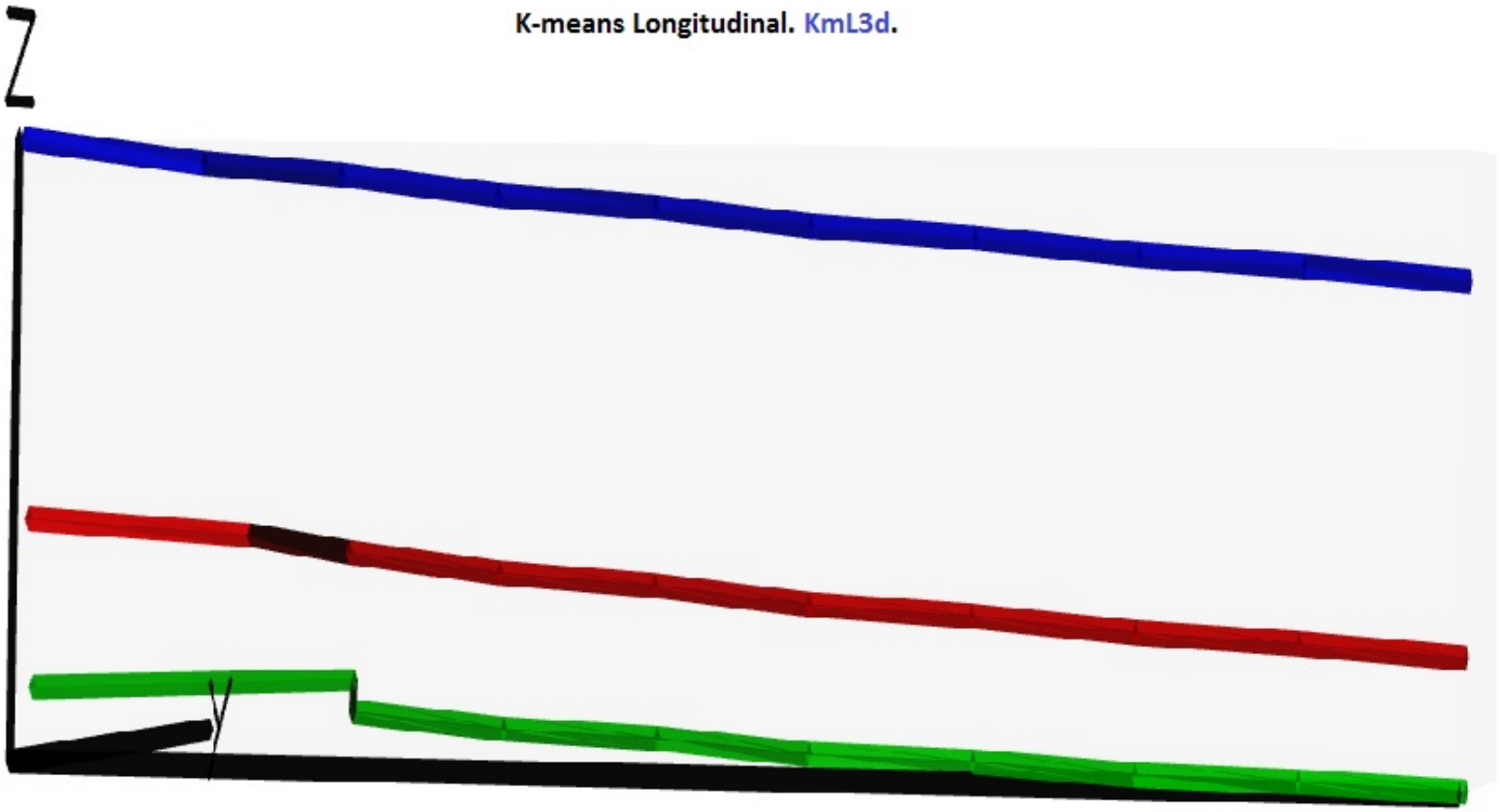


Apliación

Modelo 2. Trayectorias



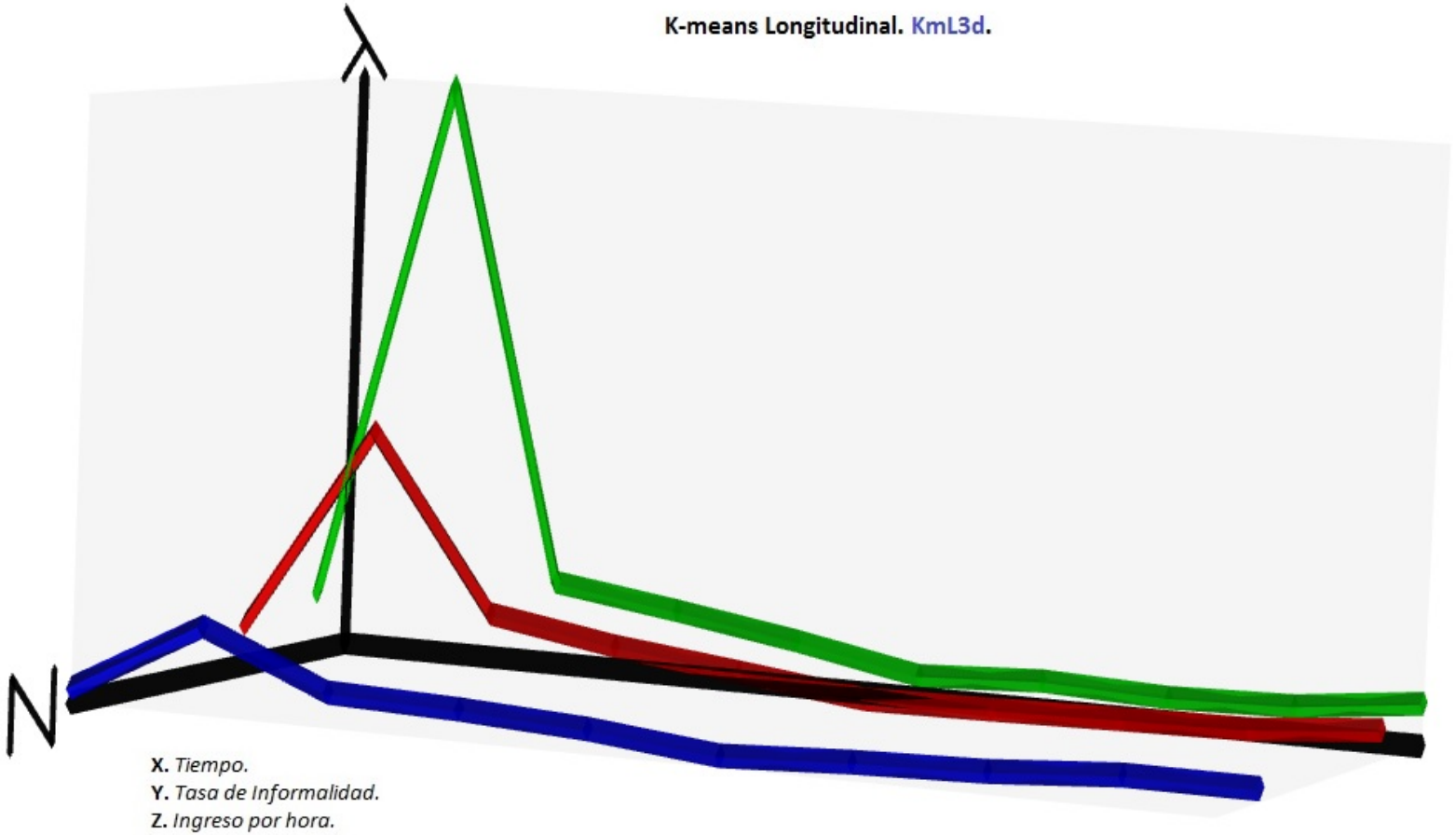

Calinski.Harabatz
Sorted
200
250
300
350
0
10
20
30
40
50
60
Rerolling



X. *Tiempo.*
Y. *Tasa de Informalidad.*
Z. *Ingreso por hora.*

Aplicación

Modelo 3. Trayectorias



Resultados

Modelo 1.

- Para la variable trayectoria estimada a partir del **modelo 1** se obtuvieron *dos grupos*.
- El Grupo "A" se compone por por las cohortes, que en el año 1989, tenían edades entre 20 y 60 años, para el caso de los varones y entre 24 y 60 para las mujeres. La trayectoria media del cluster es estable, y su tendencia corresponde los indicadores económicos y laborales.
- El Grupo "B" se compone por cohortes con niveles de informalidad relativamente altos. Se distinguen, a su vez, dos sub-grupos demográficos. El primero integra las cohortes que se integran al mercado de trabajo: varones entre 15 y 19 y mujeres entre 15 y 23 años. El segundo incluye cohortes de mujeres entre 53 y 60 (en 1989).

Resultados

Modelo 1.

GRUPO A

PSEUDO PANEL		GENERO	
Varon	Mujer	Varon	Mujer
i_v20		41	29
a		58,6	41,4
i_v23			
i_v24	i_m24	89,1	63,0
a	a		
i_v52	i_m52	n	
		70	
i_v53		78,3	
a			
i_v60			

GRUPO B

	PSEUDO PANEL		GENERO	
	Varon	Mujer	Varon	Mujer
Ingreso **V** MFT	i_v15	i_m15	5	17
	a	a	22,7	77,3
	i_v19	i_m19		
Ingreso **M** MFT		i_m20	10,9	37,0
		a		
		i_m23	n	
			22	
		i_m53	21,7	
		a		
		i_m60		

Bibliografía

- **Canavire-Bacarreza G., Urrego J. A., Saavedra F. (2017).** "Informality and Mobility in the Labor Market: A pseudo-panel's approach". Revista Latinoamericana de Desarrollo Económico. No.27. La Paz. Mayo, 2017.pp 57-75.
- **Diggle, P.J., Heagerty, P., Liang, K-Y and Zeger, S.L. (2002).** Analysis of Longitudinal Data (second edition). Oxford: Oxford University Press.
- **Garre M., Cuadrado J.J., Sicilia M.A., Rodriguez D. and Rejas R. (2007).** "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, Vol.3, No. 1, 2007.

Bibliografía

- **Guiller, M. (2017).** "Pseudo-panel methods and an example of application to Household Wealth data". Economie et Statistique, 2017, pp. 109-130.
- **Genolini C., Alacoque X., Sentenac M. and Arnaud C. (2015).** "Kml and kml3d: R packages to Cluster Longitudinal Data". Journal of Statistical Software. Volume 65, Issue 4. May 2015.
- **Meng Y., Brennan A., Purshouse R. and Otros (2014).** "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. Volume 34, March 2014, Pages 96-103.

Muchas Gracias