Metodologías de análisis longitudinal para determinar los Perfiles de la Informalidad laboral en Gran Córdoba.

Iglesias, Maximiliano Luján

Stimolo, María Inés

Instituto de Estadística y Demografía.
Facultad de Ciencias Económicas.
Universidad Nacional de Córdoba.

Junio, 2019

1. Problema.

- 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.

Canavire-Bacarreza, Urrego & Saavedra (2017).

- El presente trabajo tiene por **objetivo** el desarrollo de metodologías estadísticas que posibiliten incorporar la dimensión temporal como un factor clave para un análisis lo más completo posible de la dinámica y estructura de la problemática objeto de estudio, como así también, la relación las entre sus múltiples factores y determinantes.

1.1 Datos longitudinales.

*“La característica definitoria de un **estudio longitudinal** es que los individuos se miden repetidamente a través del tiempo.*

Los estudios longitudinales contrastan con los estudios transversales, en los que se mide un resultado único para cada individuo.

*Si bien a menudo es posible abordar las mismas preguntas científicas con un estudio longitudinal o transversal, la principal **ventaja** de los longitudinales es su capacidad para separar lo que en el contexto de los estudios de población se llama **efectos de cohorte y de la edad**.”*

Diggle P, Heagerty P, Kung-Yee, L &Scott L (2004).

2.1. Pseudo Panel.

DATOS DE PANEL.

En estadística y econometría, el término de **datos de panel** hace referencia a datos que combinan una dimensión temporal con otra transversal.

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.

$$(1)\ y_{it} = x_{it}\beta + \alpha_i + \varepsilon_{it} \quad i = 1, \ldots, N. \quad t = 1, \ldots, T.$$

y_{it} es la variable de interés.

x_{it} es el vector (lineal) de P variables explicativas.

β indica el efecto de estas variables (vector de parámetros).

α_i Efecto individual que captura todos los determinantes de la variable de interés que están fijos en el tiempo.

Guillerm, M (2017).

2.1. Pseudo Panel.

DATOS DE PANEL.

Limitaciones.

- Suelen ser paneles 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.

Perera,J (2006)

2.1. Pseudo Panel.

PSEUDO PANEL.

- La metodología que se pretende trabajar **(Deaton,1985)** tiene como objetivo, superar estas limitaciones mediante la construcción de paneles “sintéticos”.
- Esto se logra, reemplazando las observaciones individuales del panel original con medias de subgrupos de la población, es decir, subgrupos de individuos de los que se puede identificar su aparición en repetidas encuestas transversales.
- Los factores para definir subgrupos (miembros del pseudo-panel) deben ser o suponerse invariantes en el tiempo (por ejemplo, año de nacimiento, género, etnia).

- $\boldsymbol{N}$ conjunto de datos en secciones repetidas, **C** número de subgrupos definidos y $\boldsymbol{n_c}$ numero de observaciones dentro del grupo. **N = C * $\boldsymbol{n_c}$ * T.**

- $\boldsymbol{n_c}$. Trade off entre homogeneidad y robustez.

Meng Y, Brennan A, Purshouse R & Hill-McManus D (2014).

2.1. Pseudo Panel.

PSEUDO PANEL.

El **modelo de pseudo panel**, como se mencionó, se define a través de las medias de los subgrupos (cohortes) en **(1)** sobre los n_c individuos observados para la cohorte "c" en el periodo "t".

$$(2)\ \bar{y}_{ct} = \bar{x}_{ct}\beta + \bar{\alpha}_{ct} + \bar{\varepsilon}_{ct} \quad c = 1, \ldots, C. \quad t = 1, \ldots, T.$$

$\bar{y}_{ct} = E(y_{it} | \ i \in c, t).$ $\bar{y}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c} y_{it}.$

$\alpha_{c(t)}$ es el efecto fijo a nivel cohorte.

2.1. Pseudo Panel.

VENTAJAS

- Posibilita el seguimiento de cohortes a lo largo del tiempo en secciones transversales repetidas, generando series de tiempo para las medias de los subgrupos ($\bar{y}_{ct}$)que se pueden usar como si los datos del panel estuvieran disponibles.
- Atenúa en gran medida el sesgo derivado de los errores de medida $y_{it} = y_{it}^{*} + \varepsilon_{it}$.
- Cuando el número de individuos en cada cohorte es grande ($n_c \to \infty$), se tiene que $\bar{\varepsilon}_{ct} = \frac{1}{n_c}\sum_{i=1}^{n_c} \varepsilon_{it} \xrightarrow{P} E(\varepsilon_{it}) = 0$
- Un tamaño de cohorte/año suficientemente grande ($n_c > 100$) estimar consistentemente los parámetros de la ecuación y que no sean afecten por los errores de medida.
  Deaton (1985), Moffit (1993), Collado (1997), Mckenzie (2004) y Verbeek y Vella (2005)

2.2. Clustering longitudinal.

CLUSTERING TEMPORAL

Técnicas de clúster-temporal.

- Las técnicas de clúster-temporal combinan similitudes de contenido y adyacencia temporal en una sola representación. Esto implica que deben utilizarse algoritmos de agrupamiento temporal que tengan en cuenta los vecinos temporales de los objetos para extraer conocimiento útil.

- La estructura temporal se incorporará considerando un dominio de tiempo válido $D_{VT} = \{C_1^v, C_2^v, \ldots, C_k^v\}$ y el dominio del tiempo de transacción $D_{TT} = \{C_1^t, C_2^t, \ldots, C_k^t\}$, siendo $c_b = (c_v, c_t)$ los intervalos de tiempo o “cronos”.

k-means for Longitudinal data (Genolini & Falissard) constituyen una implementación de k-means diseñados para funcionar específicamente en trayectorias (kml) o en trayectorias conjuntas (kml3d).

2.2. Clustering longitudinal.

kml y kml3d

- Sea **S** un conjunto de **C** elementos (cohortes de individuos).

- Para cada cohorte "**i**", tenemos **p** variables de resultados $\mathbf{Y_{..A}}$, $\mathbf{Y_{..B}}$,. . ., $\mathbf{Y_{..P}}$ medidas en **t** diferentes tiempos.

- Se denomina a $\mathbf{Y_{..A}}$ como "variable trayectoria única" o "variable trayectoria", mientras que varias variables trayectorias consideradas conjuntamente ($\mathbf{Y_{..A}}$, $\mathbf{Y_{..B}}$,. . ., $\mathbf{Y_{..P}}$) se llamarán "variables trayectorias conjuntas".

Para la cohorte "**i**", el valor (de medición) de la variable trayectoria $\mathbf{Y_{..A}}$ en el momento **j** se denota como $\mathbf{y_{ijA}}$.

La secuencia $\mathbf{y_{ijA} = (y_{i1A}, y_{i2A}, . . ., y_{itA})}$ se denomina trayectoria única (o trayectoria) de la medición A en la cohorte "**i**".

2.2. Clustering longitudinal.

kml y kml3d

Las trayectorias conjuntas escritas en términos de matriz

$$y_{i..} = \begin{pmatrix} y_{i1A} & y_{i2A} & \cdots & y_{itA} \\ y_{i1B} & y_{i2B} & \cdots & y_{itB} \\ \vdots & \vdots & \cdots & \vdots \\ y_{i1P} & y_{i2P} & \cdots & y_{itP} \end{pmatrix}$$

$\boldsymbol{i} = c = 1, \ldots, C.$; **p**$= 1, \ldots, P$; $\boldsymbol{j} = 1, \ldots, t.$

- Las líneas son trayectorias $y_{i.A}$ = **(y_{i1A}, y_{i2A}, . . ., y_{itA})** de una sola variable.
- Si **j** es fija, la secuencia $y_{ij.}$ se denomina *estado de la cohorte* en el momento **j**.

2.2. Clustering longitudinal.

kml y kml3d

- El objetivo del agrupamiento es dividir **S** en **k** sub-grupos homogéneos.
- Más formalmente, sea **Dist** una *función distancia* y ||. || una *norma*. Para calcular la distancia **d** entre $y_{1..}$ y $y_{2..}$

2.2. Clustering longitudinal.

kml y kml3d

Método I.

Para cada **j** fija, definimos la distancia entre $y_{1..}$ y $y_{2..}$ (distancia entre el *estado de las cohortes* en el momento **j**) como

$$\boldsymbol{d_{j.}(y_{1j.}, y_{2j.}) = Dist(y_{1j.}, y_{2j.})} \quad .$$

El resultado es un "*vector de* $\boldsymbol{t}$ *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 $y_{1..}$ y $y_{2..}$ es

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

2.2. Clustering longitudinal.

kml y kml3d

Método II.

La distancia $\boldsymbol{d'}$ entre $\boldsymbol{y_{1..}}$ y $\boldsymbol{y_{2..}}$ para cada variable **X**, definimos la distancia entre $\boldsymbol{y_{1.X}}$ y $\boldsymbol{y_{2.X}}$ (distancia entre dos trayectorias individuales **X**) como

$$\boldsymbol{d_{.X}(y_{1.X}, y_{2.X}) = Dist(y_{1.X}, y_{2.X})} \quad .$$

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

$$\boldsymbol{(d_{.A}(y_{1.A}, y_{2.A}), d_{.B}(y_{1.B}, y_{2.B}), \ldots, d_{.P}(y_{1.P}, y_{2.P}))}$$

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

Finalmente, la distancia entre $y_{1..}$ y $y_{2..}$ es

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

2.2. Clustering longitudinal.

kml y kml3d

La elección de la norma $\|.\|$ y la distancia $Dist$ pueden llevar a la definición de un gran número de distancias. Por el contrario, en el caso de $\|.\|$es la p-norma (p-norm) estándar y la $Dist$ es la distancia de **Minkowsky** con los *p* parámetros, eligiendo el método d y d'da el mismo resultado $d(y_{1..}, y_{2..}) = d'(y_{1..}, y_{2..})$.

$$Dist(y_{1..}, y_{2..}) = \sqrt[p]{\sum_{j,X} \left| y_{1jX} - y_{2jX} \right|^p}$$

$p = 2$ Distancia Euclideana.

$p = 1$ Distancia Manhattan.

$p \to +\infty$ Distancia Máxima.

2.2. Clustering longitudinal.

$$d(y_{1..}, y_{2..}) = \sqrt[p]{\sum_j (d_{j.}(y_{1j.}, y_{2j.}))^p} = \sqrt[p]{\sum_j \left(\sqrt[p]{\sum_X |y_{1jX} - y_{2jX}|^p} \right)^p}$$

$$= \sqrt[p]{\sum_j \sum_X |y_{1jX} - y_{2jX}|^p} = \sqrt[p]{\sum_X \left(\sqrt[p]{\sum_j |y_{1jX} - y_{2jX}|^p} \right)^p}$$

$$= \sqrt[p]{\sum_X (d_{.X}(y_{1.X}, y_{2.X}))^p} = d'(y_{1..}, y_{2..})$$

Genolini C and Falissard B (2010).

2.2. Clustering longitudinal.

kml y kml3d

ÓPTIMO. Caliliski T & Harabasz J (1974).

$\boldsymbol{n_m}$: Número de trayectorias en el grupo $\boldsymbol{m}$.
$\boldsymbol{\bar{y}_m}$: Trayectoria media del grupo $\boldsymbol{m}$.
$\boldsymbol{\bar{y}}$: Trayectoria media de todo el conjunto de datos $\boldsymbol{S}$.
$\boldsymbol{y_{mk}}$: Trayectoria media de $\boldsymbol{k}$ en el grupo $\boldsymbol{m}$.

MATRIZ VARIANZA-ENTRE. $\boldsymbol{B} = \sum_{m=1}^{g} n_m(\bar{y}_m - \bar{y})\,(\bar{y}_m - \bar{y})'$

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

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

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

3. Aplicación.

DATOS.

Encuesta Permanente de Hogares de modalidad puntual (**EPH puntual**) con dos ondas anuales, en mayo y octubre, realizado por el Instituto Nacional de Estadísticas y Censos (INDEC) considerando el aglomerado **Gran Córdoba** entre los años **1989** y **1995**.

PSEUDO PANELES.

Se construyeron paneles “sintéticos” a partir de la base de datos R2 de la EPH de 1989 primera onda (mayo), considerando la edad simple y el genero de los sujetos.

Se incluyeron al análisis los grupos que tenían entre 15 y 60 años en mayo de 1989, es decir las cohorte nacidas entre 1929 y 1974.

INFORMALIDAD.

Se definió la condición de informalidad en los trabajadores asalariados como la denegación total o parcial de alguno de los siguientes derechos y/o beneficios: vacaciones, aguinaldo, indemnización por despido, descuento jubilatorio.

$\bar{y}_{ct}$ el promedio de informalidad (o tasa) de la cohorte C en el momento t que asume valores entre 0 y 1.

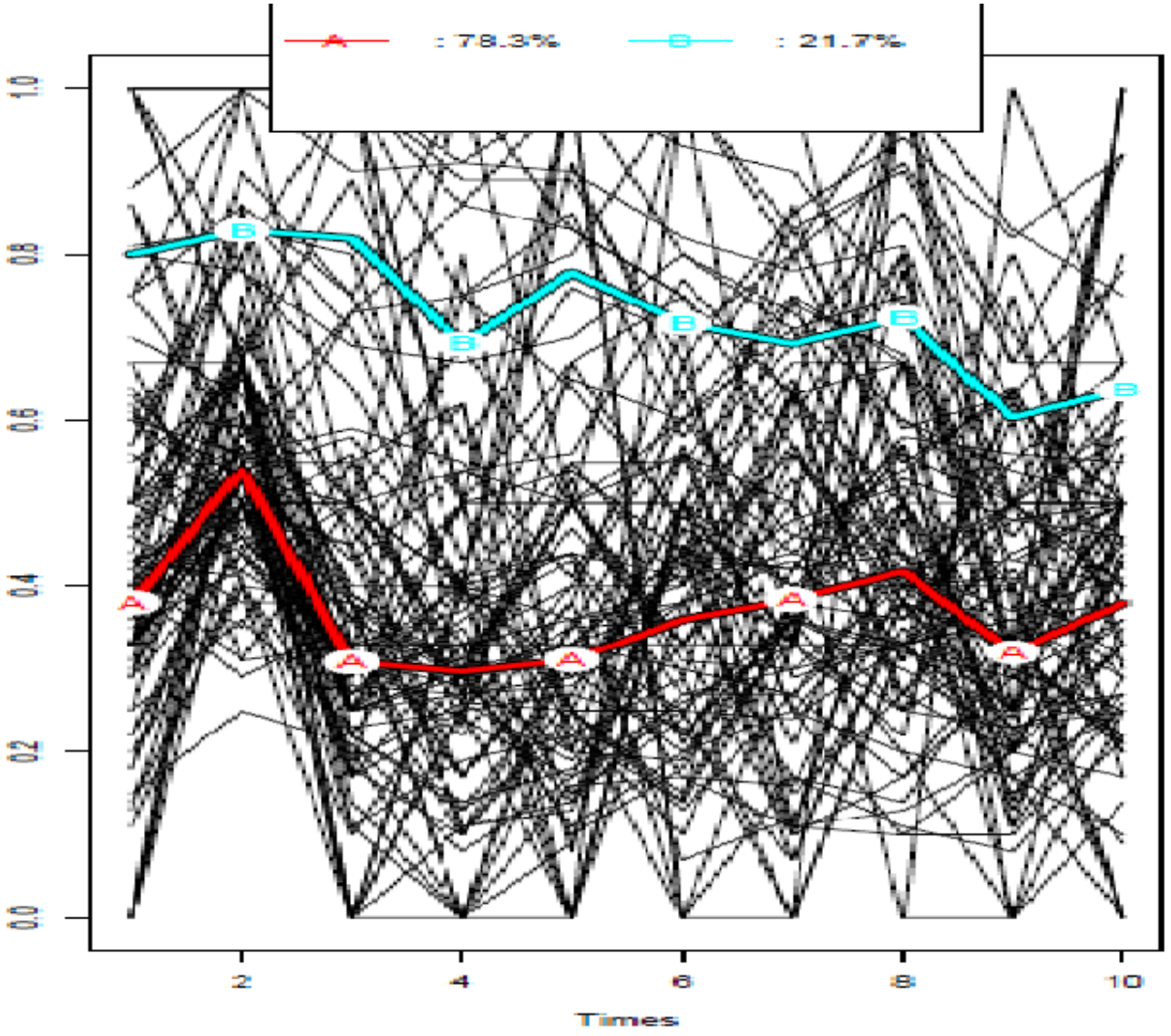
$\bar{y}_{ct} = \frac{1}{n_c} \sum_{i=1}^{n_c} y_{it}$.

3. Aplicación.

PSEUDO PANELES.

Nacimiento	Genero	CodPsPanel	1989_01	1990_01	1990_03	1991_01	1991_03	1992_01	1992_03	1993_01	1993_03	1994_0
1929	Mujer	i_m60	0,75	1	1	0,5	1	0	0	1	0,25	1
1929	Varon	i_v60	0,33	0,5	0,17	0,25	0	0,5	0,33	1		0,5
1930	Mujer	i_m59	0,6	0,8	1	0,5	0,67	0,75		0,6	0,5	0,5
1930	Varon	i_v59	0,25	0,67	0	0,4	0,2		0,5	0	0	
1931	Mujer	i_m58	0,75	0,86	0,75	1		1	1	1		
1931	Varon	i_v58	0,5	0,57	0	0	0,25	0,14	0,63	0,67	0,33	0,5
1932	Mujer	i_m57	0,33	0,5	0,5	1	0,5	1	0,33	1	0,5	0,67
1932	Varon	i_v57	0	0,67	0	0,25	0,27	0,5	0,4	0,4	0,4	0
1933	Mujer	i_m56	0,5	0,67	0	0,4	0	0,5	0,67	0,67	0	0,5
1933	Varon	i_v56	0,14	0,4	0,22	0,13	0,3	0,2	0,33	0,25	0,33	1
1934	Mujer	i_m55	1	0,67	1	0,25	1	0		0	1	0,67
1934	Varon	i_v55	0	0,6	0,25	0,42	0,25	0	0,33	0,33	0,4	0,2
1935	Mujer	i_m54	0,5	0,9	0,75	0,5	0	1	1	0,5	0,8	0,5
1935	Varon	i_v54	0,44	0,69	0,11	0,19	0,22	0,14	0,36	0,5	0,3	0,25
1936	Mujer	i_m53	1	1	0,5	0,4	1	0,8	0,71	0,75	0,5	0,8
1936	Varon	i_v53	0,33	0,5	0,38	0,17	0,33	0,22	0,36	0,3	0,22	0,33
1937	Mujer	i_m52	0,67	0,67	0,25	0,5	0,25	0,5	0,5	0,8	0,33	1
1937	Varon	i_v52	0,33	0,67	0,38	0	0,33	0,38	0,6	0,8	0,2	1
1938	Mujer	i_m51	0,6	0,56	0	0,25	0,4	0,6	0,75	0,67	0,43	0,5
1938	Varon	i_v51	0,43	0,41	0,1	0,3	0,3	0,57	0,25	0,2	0,17	0
1939	Mujer	i_m50	0,67	0,4	0,38	0	0,36	0,4	0,57	0,4	0,5	0,6
1939	Varon	i_v50	0,43	0,54	0,11	0	0,09	0,5	0,33	0,67	0,11	0,5
1940	Mujer	i_m49	0	0,5	0,33	0,33	0	0,5	0,8	1	0,38	0
1940	Varon	i_v49	0,13	0,55	0,25	0,3	0,11	0,2	0,25	0,1	0,2	0,17

3. Aplicación.



4.Referencias.

- **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. N°.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. & 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.
- **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. & Arnaud C. (2015).** “Kml and kml3d: R packages to Cluster Longitudina Data”. Journal of Statistical Software. Volume 65, Issue 4. May 2015.
- **Meng Y., Brennan A., Purshouse R. & 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.

GRACIAS POR SU TIEMPO

2.2. Clustering longitudinal.

ALGORITMO K-MEANS

¿Por qué K-means?

(1) No requiere ninguna normalidad o supuestos paramétricos dentro de los grupos (pueden ser más eficientes bajo un supuesto dado, pero no requieren uno; esto puede ser de gran interés cuando la tarea es agrupar datos sobre los cuales no hay información previa disponible).

(2) Es probable que sean más **robustos** en cuanto a convergencia numérica.

(3) En el contexto particular de datos longitudinales, **no requieren de ningún supuesto con respecto a la forma de la trayectoria.**

(4) En el contexto particular de datos longitudinales, son **independientes de la escala de tiempo**.

Genolini C and Falissard B (2010).

2.2. Clustering longitudinal.

kml y kml3d

¿Por qué kml?

(1) Proporciona margen para tratar con los **datos faltantes**.

(2) Ejecuta el algoritmo varias veces, variando las condiciones de inicio y/o el número de agrupaciones buscadas.

(3) Su **interfaz gráfica** ayuda al usuario a elegir el número apropiado de agrupaciones cuando el criterio clásico no es eficiente.

(4) Da resultados mucho mejores al Pro Traj en **trayectorias no polinómicas**.

Genolini C and Falissard B (2010).