ISSN 2013-9004 (digital); ISSN 0210-2862 (paper)
https://doi.org/10.5565/rev/papers.3079

El análisis cuantitativo de trayectorias laborales. Un estado del arte

Juan Antonio Carbonell Asíns
Centro de investigación biomédica INCLIVA, Valencia, España
jacarbonell@incliva.es

Carles Xavier Simó Noguera
Universidad de Valencia. Departamento de Sociología y Antropología Social
carles.simo@uv.es

Recepción: 18-01-2022
Aceptación: 28-07-2022
Publicación: 20-10-2022

Cita recomendada: Carbonell Asíns, Juan Antonio y Simó Noguera, Carles Xavier (2022). «El análisis cuantitativo de trayectorias laborales. Un estado del arte». *Papers*, 107 (4), e3079. <https://doi.org/10.5565/rev/papers.3079>

Resumen

La metodología cuantitativa aplicada al estudio de las trayectorias laborales ha experimentado un rápido auge que se ha extendido más allá del tradicional análisis de secuencias. El presente artículo es un estado del arte del desarrollo de nuevas técnicas estadísticas que pueden aplicarse o ya se aplican al estudio de trayectorias laborales. Además, incluimos sugerencias de *software* estadístico para la aplicación de cada una de las técnicas descritas.

A lo largo de todo el texto, podrá observarse que la descripción de cada técnica se ha realizado desde un punto de vista conceptual, con el objetivo de llegar a un público amplio, que no necesite poseer una fuerte formación estadística. Es mediante esta visión general que mostraremos las debilidades y fortalezas que cada técnica presenta, así como el hilo conductor que nos lleva de una a otra. Este trabajo parte de una perspectiva global en el estudio de las trayectorias laborales que luego tiende hacia una perspectiva más compleja, en que el interés se centra en una pequeña parte de dichas trayectorias o incluso de simples cambios de estado.

La creciente complejidad de los modelos desarrollados será objeto de discusión final debido a los nuevos retos que presentan su aplicación e implementación. Es en este contexto donde argumentaremos la necesidad de un perfil estadístico en el marco de los proyectos de investigación, tal y como sucede en otras áreas científicas. Finalmente, debatiremos la utilidad de la estadística bayesiana a la hora de enfrentarse a modelización compleja.

Palabras clave: trayectorias laborales; métodos estadísticos; análisis de secuencias; modelos de Márkov; modelos compartimentales

Abstract. *The quantitative analysis of career paths. A review of the latest techniques*

Quantitative methodology applied to the study of career paths has undergone a rapid boom that goes beyond traditional sequence analysis. This paper reviews the latest statistical techniques that can be or are already being applied to the study of career paths. We also include suggested statistical software for each of the techniques described. All techniques described here are analysed from a conceptual point of view in order to reach a broader readership who may not have a strong statistical background. It is through this overview that we will show the weaknesses and strengths of each technique, as well as a linking thread that takes us from one to another. The paper starts with a general survey of the study of career paths, before going into greater depth, focusing on a small part of a given career path or even on simple changes of status. The increasing complexity of the models described here will be the subject of a final discussion, looking at the new challenges presented by their application and implementation. It is against this background that we will argue for the need for a statistical profile within the context of research projects, as is already the case in other scientific areas. Finally, we will discuss the usefulness of Bayesian statistics in analysing complex modelling.

Keywords: career paths; statistical methods; sequence analysis; Markov models; compartment models

Sumario

Introducción

El presente artículo es una revisión del estado del arte en el estudio de las trayectorias laborales desde una perspectiva cuantitativa. El estudio cuantitativo de trayectorias laborales contempla el uso de datos longitudinales para seguir a una o varias cohortes a lo largo del tiempo. Se requiere, por tanto, la recogida de datos sobre la misma persona a lo largo de toda la ventana temporal.

El principal objetivo de este trabajo es dar una visión general y conceptual de las diferentes perspectivas metodológicas que podemos aplicar en el estudio de trayectorias laborales. Por lo tanto, este artículo no debe entenderse como una guía técnica para la aplicación de las metodologías aquí descritas. Aunque en este caso nos centremos en trayectorias laborales, todas las metodologías aquí descritas pueden aplicarse a cualquier otro tipo de trayectorias. Además, presentaremos las virtudes, debilidades y retos de cada una de ellas con el objetivo de identificar posibles vías de investigación futura.

Hemos decidido utilizar los nombres en inglés al referirnos a las técnicas, ya que su traducción puede conllevar una pérdida de significado, además, muchas

de ellas no son muy conocidas en el ámbito de la lengua castellana. Comenzaremos describiendo el *sequence analysis* como el principal método clásico para describir las trayectorias laborales. Además, discutiremos los desarrollos recientes que combinan el *sequence analysis* (SA) con el *event history analysis* (EHA). Tanto *el competing trajectory analysis* (CTA) como el *sequence analysis multistate model procedure* (SAMM) se alejan del enfoque global del SA para centrar su atención en el efecto de variables tiempo-dependientes y un subconjunto previamente definido de trayectorias en lugar de secuencias completas. El *sequence history analysis* (SHA) se centra en la relación entre la trayectoria pasada de un individuo y el evento posterior. Finalmente, discutiremos enfoques basados en modelos —conocidos como los modelos multiestado—, entre los que encontramos los modelos compartimentales y los modelos basados en estados no observados, como los modelos de Márkov con o sin perspectiva de clases latentes. Mostraremos cómo estas metodologías han evolucionado desde el estudio de trayectorias desde un punto de vista global a una perspectiva más específica y compleja.

Posteriormente, discutiremos las implicaciones de la creciente complejidad de las técnicas aquí descritas. Haremos referencia al *software* estadístico R (R Core Team, 2019), dada su capacidad para evolucionar más rápidamente comparado con otros *softwares* como Stata o SPSS. La naturaleza de R permite fácilmente crear paquetes y funciones aplicables de forma práctica al estudio de las trayectorias laborales. Otro aspecto que discutiremos se refiere al debate epistemológico que todavía no está completamente resuelto en el ámbito de la estadística. La crisis del p valor (Denworth, 2019; Wasserstein et al., 2019), la creciente complejidad de los modelos propuestos, entre otros motivos, han suscitado un amplio debate sobre las ventajas y desventajas entre las perspectivas bayesiana y frecuentista.

Finalmente, describiremos las principales conclusiones que se extraen de este artículo.

1. Análisis de secuencias

El SA es una técnica de minería de datos. Se describió por primera vez en el campo de la biología con el fin de comparar dos hebras de ADN para establecer una distancia entre ellas o, análogamente, una similitud (Kruskal, 1983). Los primeros usos del SA en sociología pueden encontrarse en Abbott (1983) y Abbott y Forrest (1986). Aunque el SA se considera una metodología cuantitativa, Andrew Abbott la utilizó desde una perspectiva cualitativa en el contexto de la sociología histórica y narrativa. Lentamente, la popularidad del SA comenzó a aumentar con el desarrollo de computadoras con procesadores más rápidos que permitieron a las investigaciones analizar secuencias más largas e individuales. Además, los paquetes estadísticos como Stata incluyeron esta metodología en su *software*, lo que hace que su uso sea más atractivo. Si por el contrario se prefiere utilizar R, el paquete que podemos recomendar es TraMineR (Gabadinho et al., 2011).

La idea principal detrás del SA es comparar secuencias individuales y establecer distancias entre ellas desde una perspectiva global. Entendemos una secuencia como la colección ordenada de los estados experimentados durante un período, típicamente observados a intervalos regulares (Piccarreta y Matthias, 2019). La distancia entre las secuencias individuales no es única, y se han propuesto varios tipos de distancia. El método principal para el cálculo de la distancia se llama *optimal matching* (OM). La idea detrás del OM es que la distancia de dos secuencias se incrementa a medida que aumenta el número de operaciones necesarias para transformar una secuencia en la otra. Estas operaciones pueden ser la inserción de un estado en una posición específica, cambiar un estado a otro o eliminar un estado específico. El resultado de este método se denomina distancia de Levenshtein (Levenshtein, 1966).

Levine (2000) y Wu (2000) señalan una limitación crucial de la OM. El coste relativo de cada una de las posibles operaciones utilizadas para establecer la distancia entre dos secuencias es definido *a priori* por el investigador o investigadora, lo que conduce a la subjetividad debido a la falta de criterios teóricos para asignar pesos a cada una de estas operaciones. Se han realizado varias modificaciones en la metodología OM original con el fin de superar sus limitaciones (Aisenbrey y Fasang, 2010; Gauthier et al., 2009). Lesnard (2010) proporciona una modificación dinámica de la distancia de Hamming que se define como el número de sustituciones necesarias para traducir una secuencia en otra. El enfoque de Lesnard pondera estas sustituciones de acuerdo con su posición. Otra limitación, a la hora de calcular la matriz de distancia entre todas las secuencias, es la presencia de datos censurados y faltantes. Este problema deriva del SA y no de cálculos de distancia, porque el SA se centra en toda la trayectoria y no en un período específico determinado. La naturaleza holística de este método es incompatible con la censura o las lagunas por falta de datos, ya que no se puede hacer distinción entre ambos, de manera que muchos criterios de disimilitud se tratarían como un nuevo estado. Halpin (2016) propone múltiples métodos de imputación con el fin de estimar los estados faltantes y también desarrolla una nueva metodología de cómo tratarlos. Sin embargo, esta sigue siendo la principal limitación del SA (Aisenbrey y Fasang, 2010).

No obstante, se han desarrollado diversas técnicas para visualizar los resultados de estas trayectorias, como el *parallel coordinate plot* o *index plot* y sus extensiones (Bürgin y Ritschard, 2014; Kohler y Brzinski-Fay, 2005; Piccarreta, 2017; Scherer, 2001).

Una vez que se calculan todas las distancias entre todos los individuos en forma de una matriz de distancias, es el momento de buscar agrupaciones mediante métodos tradicionales multivariantes como el *clustering* jerárquico. Sin embargo, el número de tipologías o clústeres aún no se ha decidido y, aunque se han descrito muchos métodos, no está claro cuál de estos es el adecuado para el estudio de las trayectorias laborales. Vale la pena mencionar que los métodos basados íntegramente en los datos (*data-driven*) pueden producir resultados que no estén en consonancia con la teoría sociológica. A pesar de la

obvia relevancia de esta materia, debemos señalar que la literatura todavía no ofrece herramientas para evaluar la validez sociológica de una tipología-clúster (Piccarreta y Matthias, 2019a). Los autores sugieren la necesidad de definir pautas y procedimientos que permitan evaluar la adecuación de los clústeres elegidos, ya que es uno de los desafíos más importantes que enfrenta no solo y sí en particular el SA.

El análisis de secuencias se califica como un enfoque holístico para identificar trayectorias típicas, pero se puede combinar con otras técnicas para describir más a fondo los factores que definen estas trayectorias. En el caso de tener más de dos clústeres, la regresión multinomial puede llevarse a cabo creando una nueva variable cuyos valores correspondan a cada una de las tipologías encontradas. La regresión multinomial tiene como objetivo el estudio de una variable con más de dos categorías y sin orden en función de otras variables. Por lo tanto, para cada individuo, la variable dependiente será el tipo de trayectoria basada en el resultado del SA, y las variables independientes serán aquellas que el investigador o investigadora consideren como factores que puedan estar asociados. En el caso de que pueda establecerse un orden lógico teóricamente sustentado entre las diferentes tipologías, puede ser más acertado el uso de una regresión ordinal, ya que su interpretación es más sencilla. Finalmente, en el caso de que dispongamos de tan solo dos tipologías, la opción será el uso de la regresión logística, la cual está pensada para variables dependientes dicotómicas. En cualquiera de los casos, el objetivo ahora será la interpretación de los coeficientes asociados a las variables independientes y su asociación con las diferentes tipologías.

Muchas han sido las aplicaciones del SA en sociología, por ejemplo, al estudio de trayectorias laborales (Brzinski-Fay, 2007; Elzinga y Liefbroer, 2007) o incluso en el contexto del mercado laboral español (López-Andreu y Verd, 2016; Verd et al., 2019). En este segundo artículo, se estudian los determinantes en las trayectorias laborales en personas jóvenes durante la recesión económica (2006-2013). Primeramente, realizan un SA y encuentran cuatro tipologías de trayectorias laborales: trayectoria laboral precaria, trayectoria de empleo temporal, trayectoria de empleo sin salario y trayectoria de empleo estable. Posteriormente, utilizan estas cuatro trayectorias en un modelo de regresión multinomial y estudian el efecto de distintas variables independientes con el objetivo de seleccionar aquellas que puedan predecir pertenecer a una u otra categoría. Verd et al. (2019) identifican diferentes factores que incrementan el riesgo de pertenecer a la trayectoria laboral precaria o a la trayectoria temporal, entre los que destaca ser inmigrante o tener un bajo nivel educativo.

1.1. Competing trajectory analysis

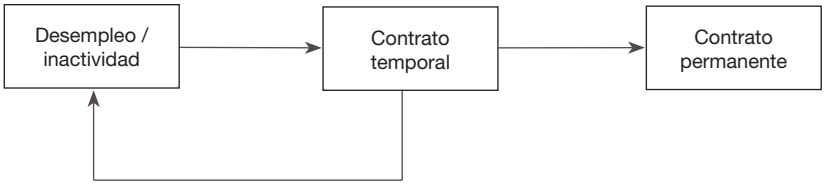
Este método se ha desarrollado en el contexto de las transiciones a la edad adulta (Studer et al., 2017), y los autores estudiaron la relación entre secularización y desempleo juvenil en las trayectorias hacia la adultez. Sus resultados mostraron que el inicio de la formación de la familia se ve pospuesto en pres-

encia de alta secularización y desempleo juvenil. Además, señalaron que el matrimonio o paternidad en edades tempranas se está volviendo menos común, mientras que hay un incremento de la cohabitación sin hijos/as.

La idea detrás de esta técnica es estudiar las trayectorias que ocurren después de la transición desde un estado inicial. En su caso, la transición de interés fue la edad adulta y detectar tipologías sobre las trayectorias. Se utilizan dos técnicas que ya estaban desarrolladas con anterioridad: el SA y los riesgos competitivos. En primer lugar, se calcularon tipologías de las distintas trayectorias a la edad adulta mediante SA, pero, en un segundo paso, se utilizó la modelización de riesgos competitivos para estimar la probabilidad de seguir una trayectoria u otra en función de las variables deseadas. Los modelos de riesgo competitivo existen en la literatura desde finales de los años setenta (Prentice et al., 1978) y tienen un uso extensivo en medicina. Este tipo de modelización está enmarcado en el análisis de supervivencia, donde el estado de interés compite con otro/s estados igualmente recurrentes. En el campo de las ciencias sociales, esta y otras técnicas están englobadas en *event history analysis* (EHA).

Pongamos como ejemplo el esquema de la figura 1, en que se parte de un estado —finalización de los estudios universitarios—, y los estados finales —desempleo, inactividad y ocupación— compiten entre sí, ya que la transición solo puede hacerse hacia uno de ellos. Aquí el objetivo es estudiar cómo afectan en el tiempo las variables independientes de interés a uno de los estados finales en términos de riesgo (*hazard ratio*: riesgo de que un suceso o estado ocurra respecto de otro). Actualmente, existen dos metodologías para estimar los riesgos: podemos usar la regresión de Cox para estimar la duración del tiempo hasta un evento censurando el resto o bien mediante la subdistribución de la función de riesgo. El primero de ellos es el más sencillo de aplicar, pues se modelizaría con la regresión de Cox, pero puede producir coeficientes sesgados al no diferenciar el motivo de la censura (Therneau et al., 2022)a standard survival curve can be thought of as a simple multi-state model with two states (alive and dead. Por ello, siempre que sea posible, debe utilizarse la segunda opción, conocida en su versión paramétrica como el modelo de Fine y Gray (Fine y Gray, 1999). Esta metodología es útil cuando la subsecuencia de interés comienza con el mismo estado —es decir, entrada en el mercado laboral— y cuando, además, la duración del primer estado es un aspecto importante de la trayectoria que hay que tener en cuenta.

Figura 1. Ejemplo de modelo de riesgos competitivos con tres estados que compiten entre sí



Fuente: elaboración propia

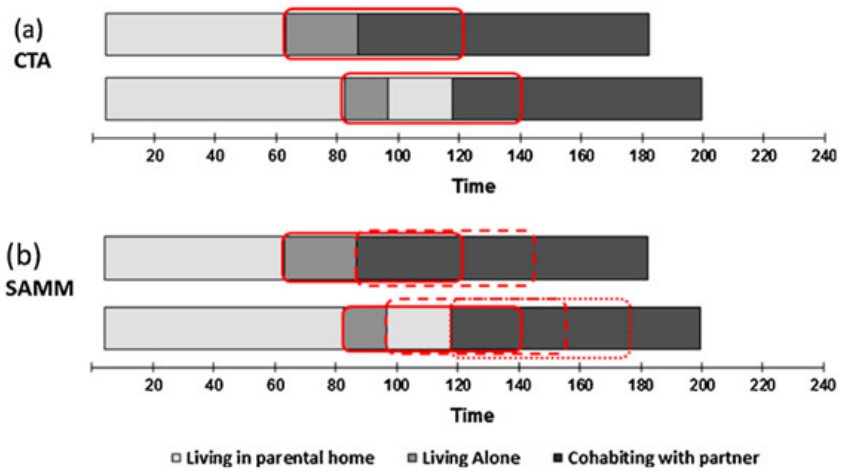
Desafortunadamente, no hay un *software* que permita realizar CTA, por lo que es necesario realizarlo en dos pasos. En un primer paso, se sugiere, al igual que se ha descrito arriba, utilizar el paquete de R TraMineR (Gabadinho et al., 2011)

1.2. Sequence analysis multistate model procedure

El método *sequence analysis multistate model procedure* (SAMM) fue desarrollado por Studer et al. (2018) en el contexto de las trayectorias laborales de las mujeres después de la reunificación alemana. Los autores incluyen un *script* con guía paso a paso para la implementación de SAMM en R.

Como muestra la figura 2, SAMM y CTA son procedimientos similares pero con ligeras diferencias. En ambos casos, la subsecuencia seleccionada tiene una longitud temporal fija, pero en CTA solo se extrae una, mientras que en SAMM se extrae más de una dependiendo del número de transiciones y de la longitud máxima permitida de la subsecuencia; dicha longitud debe definirse *a priori*. En dicha figura, observamos que, en el caso de CTA, se extrae una subsecuencia de longitud fija una vez se realiza la transición de vivir con los progenitores, que constituye el evento de interés de este ejemplo. Sin embargo, en el caso de SAMM, se seleccionan tantas subsecuencias según el número de transiciones (dos en el primer individuo y tres para el segundo) desde que se deja el hogar familiar. Al igual que en el caso de CTA, las tipologías de las subsecuencias se calculan mediante SA. En un segundo paso, se aplica un modelo para estudiar el efecto de las covariables sobre la probabilidad de iniciar cada tipo de subtrayectoria y el tiempo empleado en cada estado. Nótese que la longitud elegida de la subsecuencia es muy determinante. A medida que

Figura 2. Comparación de CTA y SAMM



Fuente: Piccarreta y Matthias (2019)

la incrementamos, incrementamos la complejidad de las interdependencias entre la primera transición y los estados que la siguen, pero si la longitud es muy larga no habrá diferencias con SA. Por lo tanto, con esta técnica podrían estudiarse efectos a corto, medio o largo plazo y sus interdependencias.

Estos métodos se alejan de la perspectiva holística, ya que se centran en las subsecuencias; los métodos SAMM y CTA pueden ser útiles para anticipar trayectorias futuras. Además, no sufren la limitación del SA y se pueden aplicar a datos censurados, lo que los hace muy útiles para algunos estudios donde la presencia de datos faltantes es inevitable.

Sin embargo, ambos métodos realmente no consideran las trayectorias pasadas. De hecho, en el caso del SAMM, solo se tiene en cuenta el estado anterior a la transición a la hora de estudiar el cambio a una determinada subsecuencia.

1.3. Sequence history analysis

Con el fin de superar las limitaciones de CTA y SAMM mencionadas anteriormente, el *sequence history analysis* (SHA) fue desarrollado por Rossignon et al. (2018). Este enfoque estudia la probabilidad de experimentar un evento no renovable o no recurrente —por ejemplo, la jubilación— por medio de un modelo discreto de EHA que incluye una variable tiempo-dependiente[1] que informa sobre las trayectorias pasadas. Intuitivamente, este método pretende estudiar un evento de interés según las trayectorias pasadas. Para ello, se establece la trayectoria de cada individuo utilizando una sucesión discreta de estados finitos y luego estos son agrupados mediante un análisis clúster. El análisis clúster se aplica para identificar subtrayectorias hasta cada punto temporal con la idea de que los individuos pueden cambiar de subtrayectoria. Dicho de otro modo, se simplifica la trayectoria laboral del individuo, de forma que ya no se tiene en cuenta cada cambio de estado sino cada cambio de subtrayectoria. Sin embargo, los autores recomiendan ajustar por edad, ya que la longitud de la trayectoria previa está muy relacionada con la edad del individuo. Finalmente, esta nueva variable tiempo-dependiente es finalmente introducida en el modelo EHA. Por lo tanto, esta metodología se desarrolla en tres fases diferenciadas: (1) construcción de las trayectorias pasadas para cada uno de los individuos, (2) aplicación de SA y (3) estimación las trayectorias pasadas en el evento de interés usando un modelo de tiempo discreto.

Los autores de esta técnica realizan una aplicación empírica al estudiar la influencia que tienen las trayectorias residenciales de la infancia en la probabilidad de emancipación. En ella observan que la llegada temprana de hijos, hijas y la monoparentalidad están asociadas a un mayor riesgo de emancipación. Además, los autores muestran que las trayectorias residenciales de la infancia

1. Una variable tiempo-dependiente es un variable cuyo contenido en un individuo puede variar con el tiempo. El ejemplo más básico sería la variable edad, ya que, después de observar a un individuo durante un año, su edad se incrementará en uno. El nivel educativo y la situación laboral también son variables tiempo-dependientes.

tienen un mayor peso que el indicador agregado de divorcio, ya que este no fue significativo en presencia de estas trayectorias.

CTA, SAMM y SHA presentan metodologías para estudiar eventos no recurrentes en los que puede ser más interesante estudiar subsecuencias que la secuencia completa. Sin embargo, dado que los métodos se basan en SA, nos encontramos con los mismos problemas de esta metodología, en especial cuando la agrupación en tipologías no es obvia (Piccarreta y Matthias, 2019). Además, la no recurrencia es una importante limitación de estos modelos que puede ser abordada por modelos multiestado como los que se describen a continuación.

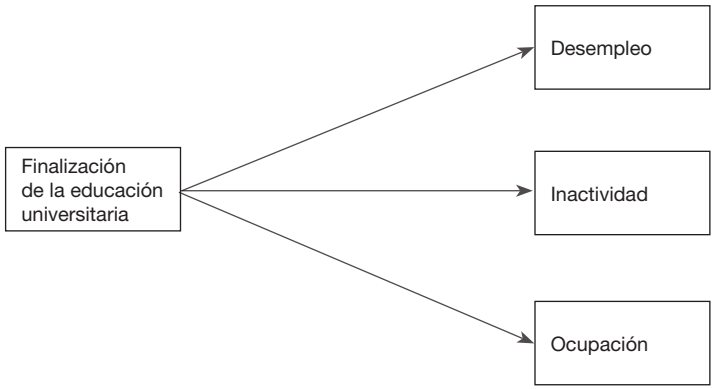
2. Modelo multiestado

Los modelos multiestado (MSM) son útiles para estudiar eventos que pueden ser recurrentes y tienen como objetivo comprender el tiempo pasado en cada estado y la probabilidad de transición de uno a otro. Más técnicamente, los MSM consisten en un proceso estocástico donde los individuos se mueven entre un conjunto de estados finitos. Los estados pueden ser absorbentes, transitorios y recurrentes. Los estados absorbentes son aquellos a partir de los cuales las personas no pueden hacer una transición a un nuevo estado, como por ejemplo los que vemos en la figura 1, donde los estados se consideran finales. Los estados transitorios son estados intermedios entre dos situaciones diferentes, mientras que los estados recurrentes o renovables son aquellos a los que los individuos pueden volver más de una vez.

La figura 2 muestra un ejemplo de un modelo multiestado con un estado transitorio —contrato temporal—, un estado absorbente —contrato permanente— y dos estados recurrentes —desempleo/inactividad y contrato temporal. Este MSM puede verse como una aplicación de un objetivo real, donde un individuo pueden estar encasillado en uno de estos tres estados con relación al mercado laboral y sería una representación práctica del dualismo del mercado laboral español (Auer y Cazes, 2000; Bustillo Llorente, 2002; Jimeno y Toharia, 1994; Simó et al., 2006). El estado de contrato permanente se representa aquí como un estado absorbente o final, y se ve representado de forma que los individuos no pueden hacer la transición a desempleados/inactivos o a contrato temporal. El estado de contrato temporal se ve como un estado transitorio para convertirse en trabajador o trabajadora con contrato permanente o, por el contrario, desempleado/a o inactivo/a. Como observación final, algunas posibles transiciones están ausentes: de desempleo/inactividad a contrato permanente, de contrato permanente a desempleo/inactividad y de contrato permanente a contrato temporal. El primero de ellos es altamente improbable, el segundo y el tercero son cada vez más posibles y de especial interés en la literatura reciente (Bachmann y Felder, 2018; Boswell y Gardner, 2018; Fallick et al., 2019).

La probabilidad de transición se basa normalmente en un proceso o cadena de Márkov. Un proceso de Márkov describe que la probabilidad de transición a un nuevo estado está totalmente definida por el estado actual. Sin embargo,

Figura 3. Ejemplo de modelo multiestado con tres estados diferentes



Fuente: elaboración propia

existen muchas variaciones de una cadena de Márkov, como que el estado futuro no solo se basa en el estado presente sino en los *m*-estados pasados. Este proceso de Márkov se conoce como un proceso de Márkov con memoria o una cadena de Márkov de orden *m*. Los siguientes métodos se basan en la cadena de Márkov de primer orden.

En el caso concreto de que los datos de los que dispongamos observen a los individuos en intervalos regulares, podremos hacer uso de modelos multiestado para un tiempo discreto. Una metodología para este tipo de dato es conocida como *state change model* (SCM) (Bonetti et al., 2013). Es un modelo paramétrico para el análisis de secuencias de tiempo discretas de un número finito de estados. Los autores describen que el principal interés de este modelo es estudiar el tiempo hasta la próxima transición de un estado en particular teniendo en cuenta la información pasada. Por lo tanto, la probabilidad condicional de que ocurra una transición se calcula en función de las variables, del tiempo en el estado anterior y de la edad en el momento de la transición a través de la regresión multinomial. Una importante limitación es la imposibilidad de estudiar las transiciones de un estado a un mismo estado, ya que este modelo de probabilidad condicional se basa en la transición de un estado pasado diferente.

2.1. Modelos compartimentales

Los modelos compartimentales son herramientas matemáticas que se han utilizado principalmente en el diseño de enfermedades infecciosas. Estos modelos matemáticos sirven para explicar cómo se comportará un objeto o sistema de objetos (Keeling y Rohani, 2008). En epidemiología, estos modelos sirven para predecir la dinámica de una epidemia a partir del conocimiento sobre la misma. A la hora de formular un modelo, hay que tener en cuenta tres aspectos relacionados que compiten entre sí: precisión, flexibilidad y transparencia.

La precisión se refiere a la posibilidad de reproducir los datos observados y su poder de predicción. Aunque siempre es deseable tener una alta precisión, hay inconvenientes, ya que aumentar la precisión generalmente implica una mayor complejidad, es decir, el proceso de incluir nuevos factores, parámetros, en el modelo matemático. Por ejemplo, volviendo a la figura 3, podríamos incluir más estados con el objetivo de aumentar la precisión. Sin embargo, uno de los problemas que surgen de tener una alta complejidad en un modelo es la gran potencia computacional necesaria para estimar todos los parámetros involucrados. Además, cuando tenemos modelos excesivamente complejos, podemos caer en la falta de transparencia, es decir, en la pérdida de capacidad de entender numérica o analíticamente cómo se interrelacionan los componentes del modelo. Normalmente, reducir la complejidad aumenta la flexibilidad, pero, por otro lado, se pierde precisión. Finalmente, la flexibilidad de un modelo se refiere a la facilidad que tiene para adaptarse a una nueva situación. Esta flexibilidad es importante en modelos que tratan de explicar dinámicas que son muy cambiantes, como las del mercado laboral.

Estos modelos matemáticos son capaces de lograr ciertos objetivos, pero son incapaces de conseguir otros. Primero describiremos sus virtudes. La predicción es una de las grandes ventajas de estos modelos, pero, como se mencionó anteriormente, es necesaria una alta precisión para ello. La predicción permite el desarrollo de políticas que anticipen un evento que podría ser catastrófico para una población: si tenemos varios modelos diferentes que nos apuntan hacia la aparición de una epidemia, podríamos tomar las medidas adecuadas para reducir su riesgo al mínimo, por ejemplo lanzando una campaña de vacunación. Sin embargo, incluso si se tienen modelos predictivos precisos, puede suceder que haya algunas áreas en las que el modelo no sea capaz de predecir correctamente, lo que implicaría que se pueden necesitar medidas específicas para detectar este comportamiento. En este contexto, el modelo propuesto en la figura 3 puede mejorarse con nuevos factores que pueden ser importantes y que aún no se habían contemplado.

Aunque estos modelos se han llevado a cabo principalmente en la disciplina de la epidemiología, cabe señalar que pueden ser aplicables a otros campos. La aplicación de este tipo de modelización a otros campos de la ciencia es uno de los mayores atractivos de estos métodos ya que casi no hay referencias de su uso en ciencias sociales. Es el caso de Santonja et al. (2008), que, en el contexto de la ciencia política, estudiaron la presión de las ideologías radicales en la sociedad española. Sin embargo, hasta donde sabemos, no hay evidencia de que se hayan utilizado modelos compartimentales en el estudio de la dinámica del mercado laboral.

Estos modelos también tienen limitaciones, pues es imposible lograr un modelo completamente preciso porque, tanto en el caso de la dinámica del mercado laboral como en el resto de estudios, siempre hay algún aspecto desconocido que no incluimos. Hay factores que no podemos controlar porque son demasiado complejos o desconocidos, ya que solo se dispone de una muestra del conjunto de la población.

El primer paso requerido para la estimación de los parámetros de interés es resolver el sistema de ecuaciones diferenciales que surgen de las relaciones entre los compartimientos. Hasta la fecha, desconocemos si existe algún paquete en R capaz de resolver el sistema de ecuaciones, pero hay programas matemáticos como Matlab (Matlab, 2019) que son capaces de hacerlo. Una vez resuelto, existen varios métodos que nos permiten estimar parámetros de un modelo compartimental. El método de mínimos cuadrados (LS) ha sido el principal de ellos. La idea detrás de este método es encontrar parámetros que produzcan valores predichos lo más cercanos posible a los datos observados. En la regresión lineal, esta estimación es directa, pero en los modelos compartimentales tenemos ecuaciones no lineales que necesitan algoritmos iterativos más complejos. Algunos de los algoritmos para la resolución de mínimos cuadrados en modelos no lineales son el Gauss-Newton, el Golub-Pereyra, para modelos de mínimos cuadrados parciales, y el PORT, propuesto por Gay (1990). Este método ha demostrado ser más útil que el resto porque permite restricciones en el rango de valores posibles de los parámetros. Todos estos algoritmos están disponibles en R, sin necesitar la instalación de ningún paquete, y son parte de la función *nls*.

El uso del método de mínimos cuadrados tiene ciertas desventajas. Este método supone que en cada momento la variación estocástica es siempre la misma y que se distribuye normalmente. Sin embargo, esto no es necesariamente cierto en nuestros datos. Es decir, si no se cumplen los supuestos de normalidad u homocedasticidad constante, los resultados obtenidos por el método de mínimos cuadrados podrían dar resultados sesgados. También se supone que las observaciones en cada momento son independientes, pero esto no suele ser cierto en el caso de datos longitudinales como estos, ya que las observaciones del mes siguiente están correlacionadas con las anteriores.

Otra forma para estimar los parámetros es a través del llamado método de máxima verosimilitud (MV). Su objetivo es maximizar la probabilidad de observar los datos dado un modelo de predicción. Por ejemplo, podemos tener un modelo que predice que la proporción de ocupados debe ser de 0,4, pero nuestros datos indican que esta proporción es de 0,8. Así, tendríamos que la probabilidad (verosimilitud) de observar que la proporción 0,8 es muy baja si la esperada es de 0,4. Sin embargo, la probabilidad de observar 0,8, si el valor esperado es 0,85, es mayor (por tanto, más verosímil) que en el caso anterior. En este caso, la función de R recomendada es *mle2* de la librería *bbmle* (Bolker y Team, 2017). Esta función permite estimar parámetros mediante el método de máxima verosimilitud. En ella podemos encontrar diversos métodos de estimación; por defecto, el método propuesto por Nelder y Mead (1965). Este método minimiza una función con n variables mediante la comparación de valores de la función en $(n + 1)$ vértices de un simplex general. Se continúa con el reemplazamiento del vértice con el valor más alto por otro punto. El simplex se adapta a la superficie local y converge hacia el mínimo final.

En nuestra experiencia no hemos obtenido buenos resultados utilizando máxima verosimilitud frecuentista o mínimos cuadrados. Sin embargo, sí obtu-

vimos resultados satisfactorios utilizando la perspectiva bayesiana, debido a la posibilidad de incluir información previa a los parámetros mediante la definición de la distribución *a priori*. Desafortunadamente, no existen paquetes para la estimación de parámetros en el contexto de modelos no lineales bayesianos, y debe ser el usuario o usuaria quien programe y especifique el modelo. Sin embargo, podemos recomendar el uso del *software* JAGS (Plummer et al., 2003), que implementa los recursos necesarios para simular la distribución *a posteriori*.

2.2. Análisis de clases latentes

En primer lugar, debemos introducir lo que se entiende por latente. Las variables latentes surgen principalmente en las ciencias sociales, ya que es aquí donde utilizamos constructos en vez de variables medibles y observables como en las ciencias experimentales (Bartholomew, 2001). Los siguientes modelos se basan en este principio que postula que existe una estructura latente subyacente a las secuencias observadas. Esta estructura debe identificar las principales características de las trayectorias filtrando la variabilidad individual que se debe a la relación probabilística entre los estados latente y observado. De una manera más intuitiva y en el contexto de las trayectorias vitales, se puede describir una estructura latente como «los planes y/o decisiones tomadas en diferentes etapas de la vida, resultando en la experiencia de estados específicos observados» (Billari y Piccarreta, 2005).

El análisis de clases latentes (Lazarsfeld y Henry, 1968) se utiliza para identificar patrones típicos —homogéneos— en las trayectorias vitales llamadas clases. Estas clases no son observables y por ello consideradas latentes, pero la pertenencia a dicha clase se puede inferir mediante datos observados a través, normalmente, de una función de máxima verosimilitud (Barban y Billari, 2012). Una importante limitación de esta técnica es que trata las mediciones de una misma variable a través de los períodos de tiempo como independientes. Esto implica que el análisis de clase latente no tiene en cuenta la correlación entre las variables dependientes del tiempo. A continuación, mostramos dos técnicas estadísticas que están englobadas en la modelización multiestado con componente de estados latentes (*hidden Markov model*) o clases latentes (*mixture hidden Markov model*). Desafortunadamente, aunque sea posible, no hemos encontrado todavía aplicaciones de estos modelos al estudio de trayectorias laborales pero sí al de trayectorias vitales.

2.2.1. Hidden Markov model

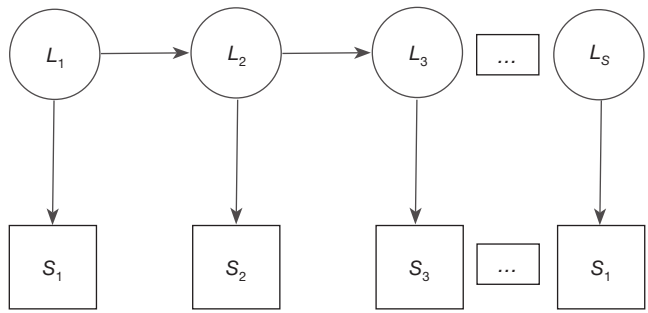
La idea principal en los *hidden Markov models* (HMM) es que el estado actual está guiado por una variable latente que sigue un proceso de Márkov. Por lo tanto, en lugar de modelizar el proceso estocástico de la variable de interés, podría ser más realista suponer que existe un proceso subyacente, es decir, latente, que modula los estados observados (Bolano et al., 2016). Esta variable latente categórica puede cambiar con el tiempo siguiendo un proceso de

Márkov de primer orden. Un ejemplo de variable latente pueden ser las relaciones que se establecen en el seno de la pareja y que no están recogidas en una encuesta, pero que afectan en las decisiones que las personas toman a la hora de elegir un trabajo u otro. Cada uno de los posibles estados de esta variable latente se conoce como estado oculto. La figura 4 muestra una representación gráfica del *hidden Markov model*. S_1 denota el primer estado observado para un individuo dado, seguido de S_2 como el segundo estado observado y S_T como el último estado observado. L_1 es el primer estado no observado, es decir, latente, para un individuo dado, seguido de L_2 como el segundo estado latente y L_T como el último estado latente. Las flechas entre estados no observados denotan la probabilidad de transición a un nuevo estado condicional al estado presente en un momento *t* . Estamos, por tanto, hablando de un proceso de Márkov de primer orden, ya que tenemos una sucesión de estados no observados cuya probabilidad de transición depende solo del estado actual. Las flechas que conectan estados latentes con estados observados indican que la probabilidad de observar un estado *S* en el tiempo *t* viene condicionada por la probabilidad de emisión del estado latente en ese mismo momento *t*, es decir, .

Pocas son las aplicaciones al estudio de trayectorias laborales, probablemente porque es una técnica relativamente nueva y su aplicación es compleja y requiere una fuerte formación en métodos estadísticos. Podemos encontrar un ejemplo en el estudio de trayectorias a la adultez en Han et al. (2016). Aquí proponen un modelo de cuatro estados para individuos de nacionalidad francesa nacidos entre 1956 y 1965. Los principales resultados que obtienen los autores se refieren a las diferencias en transición de acuerdo con el sexo y la educación. Según ellos, las mujeres de nivel educativo alto tienen mayor probabilidad de emanciparse en comparación con mujeres con menor nivel educativo, pero también posponen la formación de una familia.

El paquete de R *depmixS4* (Visser y Speekenbrink, 2010) permite la implementación de los *hidden Markov models*, así como de los *mixture hidden Markov models*, que describiremos a continuación. Además, también recomendamos *seqHMM* (Helske y Helske, 2019) para ambos tipos de modelos, pues

Figura 4. Representación gráfica del hidden Markov model



Fuente: elaboración propia adaptación de Piccarreta y Matthias (2019)

fue desarrollado por quienes, a su vez, fueron pioneras en la aplicación de este tipo de modelos al estudio de trayectorias. Sin embargo, si nuestro interés es aplicar las nuevas generalizaciones que han aparecido en este tipo de modelos como la inclusión de múltiples variables latentes o representaciones multiescala, entonces deberemos cambiar de paradigma y buscar solución en la inferencia bayesiana (Ghahramani, 2001). Además, tal y como describe Scott (2011), los métodos frecuentistas pueden traer complicaciones a la hora de tener en cuenta la incertidumbre en los parámetros a estimar. Como comentaremos en la discusión, la estadística bayesiana permite modelos más complejos donde la estimación mediante estadística frecuentista puede ser inviable.

2.2.2. Mixture hidden Markov models

Como se muestra en la figura anterior, no hay clases latentes, lo que significa que la heterogeneidad no observada puede ignorarse. Sin embargo, los *mixture hidden Markov models* (MHMM) no asumen esto, y los individuos se dividen en grupos y las secuencias dentro de cada clase provienen de HMM específicos a dicha clase (Helske et al., 2016). Por lo tanto, MHMM es una generalización de HMM donde el número de clases puede ser mayor que uno. Este modelo asume que una población puede agruparse en grupos (clases latentes) con patrones similares. MHMM considera la posibilidad de variar los submodelos para cada clúster, pero no se permiten las transiciones entre clústeres. Este método fue propuesto por primera vez por Pol y Langeheine (1990), y más tarde desarrollado por Vermunt (2008) para incluir variables constantes y variables tiempo-dependientes. Este último autor fue el encargado de nombrar el método como MHMM.

Todos los métodos multiestado que utilizan el proceso de Márkov de primer orden asumen que toda la información pasada se resume en el último estado visitado. Esta suposición puede ser su gran debilidad, ya que puede no suceder que el anterior estado resuma adecuadamente todo lo anterior en el pasado. Esta limitación se ha abordado desde dos perspectivas: incluyendo variables que informan del pasado o del presente, así como la ampliación del proceso de Márkov de un orden superior en el que se tienen en cuenta un mayor número de estados pasados. Estos métodos aumentan rápidamente su complejidad cuando crece el número de posibles estados o transiciones.

Tal y como se comentó anteriormente, tanto el paquete de R *depmixS4* (Visser y Speekenbrink, 2010) como *seqHMM* (Helske y Helske, 2019) permiten la implementación de los *mixture hidden Markov models*.

3. Discusión

El estudio de la realidad social desde el punto de vista cuantitativo viene acompañado de una creciente complejidad metodológica que requiere profesionales con una fuerte formación estadística y matemática. Las técnicas basadas en SA pueden encontrarse implementadas en Stata, pero, a la hora de escribir este artículo, los modelos multiestado aquí descritos solo lo están en R. El

software estadístico R (R Core Team, 2019) está cada vez más presente en la investigación social cuantitativa y se presenta como una alternativa al resto de programas de pago tradicionales. Su característica fundamental es que se trata de un programa libre y en constante evolución, lo que permite que cualquier persona pueda crear o compartir sus desarrollos.

Otro debate que suscitan los modelos complejos es su dificultad de implementación mediante el enfoque tradicional frecuentista y el consiguiente auge del enfoque bayesiano. En la inferencia estadística, se pueden encontrar dos perspectivas claramente diferentes: frecuentista y bayesiana. Para ilustrar la diferencia entre estos dos enfoques, digamos que nos interesa conocer el valor de un parámetro *a* de nuestro modelo. Después de la estimación a través de ML, obtendremos también sus respectivos intervalos de confianza. Es muy común caer en el error de considerar que estas estimaciones tienen una probabilidad de estar dentro del intervalo, pero esto no es correcto. En la teoría clásica (frecuentista), el parámetro *a* se consideran fijo y, normalmente, desconocido, por tanto, no tiene sentido asignar probabilidades a los parámetros. La confianza del 95 % se aplica al intervalo, no al parámetro, y lo correcto es afirmar que existe un 95 % de probabilidad de que el intervalo contenga el valor verdadero del parámetro. Sin embargo, en la estadística bayesiana, el parámetro de nuestro modelo se considera aleatorio, y expresamos su incertidumbre antes de considerar los datos en términos de probabilidad asignando *a priori* una distribución previa a ese parámetro. Posteriormente, mediante el teorema de Bayes, obtenemos la distribución *a posteriori* de *a* utilizando la distribución previa del parámetro y la verosimilitud de los datos observados.

La estadística bayesiana ya tiene muchos años de historia, pero no fue hasta finales de los años ochenta cuando comenzó a tomarse como una alternativa práctica a la teoría clásica. Es a principios de este siglo cuando ha experimentado su mayor crecimiento, coincidiendo con el desarrollo de nuevos métodos de estimación de distribuciones *a posteriori* y la gran potencia computacional de los nuevos ordenadores personales. Estos métodos han evolucionado mucho en los últimos años, desde el primero de ellos, con el (re)descubrimiento de los métodos de Montecarlo utilizando cadenas de Márkov (MCMC) (Gelfand et al., 1990; Gelfand y Smith, 1990), hasta la aproximación integrada de Laplace (INLA) propuesta por Rue et al. (2009). Otro de los motivos importantes para el crecimiento de la estadística bayesiana es la aparición de *softwares* como WinBugs (Lunn et al., 2000) o JAGS (Plummer et al., 2003), que permitieron la generación de muestras de la distribución *a posteriori* de los parámetros a estimar. Además, ahora existen paquetes para ser usados en el *software* estadístico R que implementan todo lo necesario para su utilización, como por ejemplo R2JAGS (Su y Yajima, 2020) o R2WinBugs (Sturtz et al., 2005).

Es razonable pensar que nos encontramos ante modelos cada vez más complejos tanto de implementar como de interpretar, así pues debemos recordar el aforismo de George Cox: «Todos los modelos son equivocados, pero algunos son útiles» (Box, 1976). Efectivamente, nunca encontraremos el modelo correcto capaz de explicar y predecir a la perfección, pero, ante la imposibilidad

de conseguirlo, estamos buscando uno que sea lo suficientemente útil para nosotros. Vale la pena señalar que el aumento de la complejidad no significa que un modelo sea mejor desde un punto de vista sociológico. Siempre hay una necesidad de compromiso entre la teoría social y el enfoque estadístico, pues los modelos de mayor complejidad tienden a complicar en exceso la comprensión sociológica (Piccarreta y Matthias, 2019).

4. Conclusión

Hemos presentado la evolución de estas metodologías de estudio de trayectorias desde un punto de vista holístico hasta una perspectiva más micro. Comenzamos explicando el *sequence analysis*, que parte de una visión global de la trayectoria laboral y tiene como objetivo crear agrupaciones de trayectorias similares entre sí. Continuamos con técnicas como CTA o SAMM, que centran el interés en una etapa específica de la trayectoria laboral o, como definimos anteriormente, una subsecuencia. Es en el CTA donde se introduce la idea de modelos multiestado dentro del contexto de riesgos competitivos y el intento de tener en cuenta el efecto de los estados anteriores. Este interés en tener en cuenta los estados pasados responde a que es conocido que las acciones de los individuos están inmersas en comportamientos y decisiones que ocurren durante el tiempo. Estas acciones se forman, por tanto, en el contexto de las vivencias pasadas. Es el afán de entender el efecto de la «sombra del pasado» (Bernardi et al., 2019) el que conduce a modelos mucho más complejos basados en cadenas de Márkov y variables latentes como HMM o su generalización MHMM. Un resumen de las diferentes técnicas aquí descritas se puede encontrar en Piccarreta y Matthias (2019).

Referencias bibliográficas

ABBOTT, A. (1983). «Sequences of social events: Concepts and methods for the analysis of order in social processes». *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 16 (4), 129-147. <https://doi.org/10.1080/01615440.1983.10594107>

ABBOTT, A. y FORREST, J. (1986). «Optimal matching methods for historical sequences». *The Journal of Interdisciplinary History*, 16 (3), 471-494. <https://doi.org/10.2307/204500>

AISENBREY, S. y FASANG, A. E. (2010). «New Life for Old Ideas: The "Second Wave" of Sequence Analysis Bringing the "Course" Back Into the Life Course. *Sociological Methods & Research*, 38 (3), 420-462. <https://doi.org/10.1177/0049124109357532>

AUER, P. y CAZES, S. (2000). «The resilience of the Long-Term Employment Relationship: Evidence from the Industrialized Countries». *International Labour Review*, 139 (4), 379-408. <https://doi.org/10.1111/j.1564-913X.2000.tb00525.x>

BACHMANN, R. y FELDER, R. (2018). «Job stability in Europe over the cycle». *International Labour Review*, 157 (3), 481-518. <https://doi.org/10.1111/ilr.12117>

Barban, N. y Billari, F. C. (2012). «Classifying life course trajectories: a comparison of latent class and sequence analysis». *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 61 (5), 765-784. https://doi.org/10.1111/j.1467-9876.2012.01047.x

Bartholomew, D. J. (2001). «Factor analysis and latent structure: Overview». In: Smelser, Neil J., Baltes Paul B. *International Encyclopedia of the Social & Behavioral Sciences*, Pergamon, 5249-5254. <https://doi.org/10.1016/B0-08-043076-7/00425-3>

Ben Bolker and R Development Core Team (2022). «bbmle: Tools for General Maximum Likelihood Estimation. R package version 1.0.25».

Bernardi, L.; Huinink, J. y Settersten, R. A. (2019). «The life course cube: A tool for studying lives». *Advances in Life Course Research*, 41, 100258. <https://doi.org/10.1016/J.ALCR.2018.11.004>

Billari, F. C. y Piccarreta, R. (2005). «Analyzing demographic life courses through sequence analysis». *Mathematical Population Studies*, 12 (2), 81-106. <https://doi.org/10.1080/08898480590932287>

Bolano, D.; Berchtold, A. y Ritschard, G. (2016). «A discussion on hidden Markov models for life course data». *Sequence Analysis and Related Methods (LaCOSA II)*, 241.

Bonetti, M.; Piccarreta, R. y Salford, G. (2013). «Parametric and nonparametric analysis of life courses: An application to family formation patterns». *Demography*, 50 (3), 881-902. <https://doi.org/10.1007/s13524-012-0191-z>

Boswell, W. R. y Gardner, R. G. (2018). «Employed Job Seekers and Job-to-Job Search». *The Oxford Handbook of Job Loss and Job Search*, 401.

Box, G. E. P. (1976). «Science and statistics». *Journal of the American Statistical Association*, 71 (356), 791-799. <https://doi.org/10.1080/01621459.1976.10480949>

Brzinsky-Fay, C. (2007). «Lost in transition? Labour market entry sequences of school leavers in Europe». *European Sociological Review*, 23 (4), 409-422. <https://doi.org/10.1093/esr/jcm011>

Bürgin, R. y Ritschard, G. (2014). «A decorated parallel coordinate plot for categorical longitudinal data». *The American Statistician*, 68 (2), 98-103. <https://doi.org/10.1080/00031305.2014.887591>

Bustillo Llorente, R. M. de (2002). «Mercado de trabajo y exclusión social». *Acciones e Investigaciones Sociales*, 16, 89-124. <https://doi.org/10.26754/ojs_ais/ais.200216236>

Denworth, L. (2019). «A significant problem». *Scientific American*, 321 (4), 62-67.

Elzinga, C. H. y Liefbroer, A. C. (2007). «De-standardization of family-life trajectories of young adults: A cross-national comparison using sequence analysis». *European Journal of Population/Revue Européenne de Démographie*, 23 (3-4), 225-250. <https://doi.org/10.1007/s10680-007-9133-7>

Fallick, Bruce; Haltiwanger, John; McEntarfer, Erika and Staiger, Matthew (2019). «Job-to-Job Flows and the Consequences of Job Separations». *Federal Reserve Bank of Cleveland, Working Paper* no. 19-27. <https://doi.org/10.26509/frbc-wp-201927>

Fine, J. P. y Gray, R. J. (1999). «A Proportional Hazards Model for the Subdistribution of a Competing Risk». *Journal of the American Statistical Association*, 94 (446), 496-509. <https://doi.org/10.1080/01621459.1999.10474144>

GABADINHO, A.; RITSCHARD, G.; MÜLLER, N. S. y STUDER, M. (2011). «Analyzing and Visualizing State Sequences in R with TraMineR». *Journal of Statistical Software*, 40 (4), 1-37. <https://doi.org/10.18637/jss.v040.i04>

GHAHRAMANI, Z. (2001). «An introduction to hidden Markov models and Bayesian networks». In: BUNKE, Horst and CAELLI, Terry. *Hidden Markov models: applications in computer vision*, 9-41. <https://doi.org/10.1142/9789812797605_0002>

GAUTHIER, J.-A.; WIDMER, E. D.; BUCHER, P. y NOTREDAME, C. (2009). «How much does it cost? Optimization of costs in sequence analysis of social science data». *Sociological Methods & Research*, 38 (1), 197-231. <https://doi.org/10.1177/0049124109342065>

GAY, D. M. (1990). «Usage summary for selected optimization routines». *Computing Science Technical Report*, 153 (153), 1-21.

GELFAND, A. E. y SMITH, A. F. M. (1990). «Sampling-Based Approaches to Calculating Marginal Densities». *Journal of the American Statistical Association*, 85 (410), 398-409. <https://doi.org/10.1080/01621459.1990.10476213>

GELFAND, A. E.; HILLS, S. E.; RACINE-POON, A. y SMITH, A. F. M. (1990). «Illustration of Bayesian Inference in Normal Data Models Using Gibbs Sampling». *Journal of the American Statistical Association*, 85 (412), 972-985. <https://doi.org/10.1080/01621459.1990.10474968>

HALPIN, B. (2016). «Missingness and truncation in sequence data: A non-self-identical missing state». *Sequence Analysis and Related Methods (LaCOSA II)*, 443.

HAN, Y.; LIEFBROER, A. C. y ELZINGA, C. H. (2016). «Understanding social-class differences in the transition to adulthood using Markov chain models». En: *Proceedings of the international conference on sequence analysis and related methods*, 155-177.

HELSKE, S. y HELSKE, J. (2019). «Mixture Hidden Markov Models for Sequence Data: The seqHMM Package in R». *Journal of Statistical Software*, 88 (3), 1-32. <https://doi.org/10.18637/jss.v088.i03>

HELSKE, S.; HELSKE, J. y EEROLA, M. (2016). «Analysing complex life sequence data with hidden Markov modelling». *LaCOSA II: Proceedings of the International Conference on Sequence Analysis and Related Methods*.

JIMENO, J. y TOHARIA, L. (1994). *Unemployment and labour market flexibility: Spain*. International Labour Organization.

KEELING, M. J., & ROHANI, P. (2008). *Modeling Infectious Diseases in Humans and Animals*. Princeton University Press. <https://doi.org/10.2307/j.ctvcm4gk0>

KOHLER, U. y BRZINSKY-FAY, C. (2005). «Stata tip 25: sequence index plots». *Stata Journal*, 5 (199-2016-2533), 601-602. <https://doi.org/10.1177/1536867X0500500410>

KRUSKAL, J. B. (1983). «An overview of sequence comparison: Time warps, string edits, and macromolecules». *SIAM Review*, 25 (2), 201-237. <https://doi.org/10.1137/1025045>

LAZARSFELD, P. F. y HENRY, N. W. (1968). *Latent structure analysis*. Houghton Mifflin.

LESNARD, L. (2010). «Setting cost in optimal matching to uncover contemporaneous socio-temporal patterns». *Sociological Methods & Research*, 38 (3), 389-419. <https://doi.org/10.1177/0049124110362526>

LEVENSHTEIN, V. I. (1966). «Binary codes capable of correcting deletions, insertions, and reversals». *Soviet Physics Doklady*, 10 (8), 707-710.

LEVINE, J. H. (2000). «But what have you done for us lately? Commentary on Abbott and Tsay». *Sociological Methods & Research*, 29 (1), 34-40. <https://doi.org/10.1177/0049124100029001002>

LÓPEZ-ANDREU, M. y VERD, J. M. (2016). «Employment instability and economic crisis in Spain: what are the elements that make a difference in the trajectories of younger adults?». *European Societies*, 18 (4), 315-335. <https://doi.org/10.1080/14616696.2016.1207791>

LUNN, D. J.; THOMAS, A.; BEST, N. y SPIEGELHALTER, D. (2000). «WinBUGS – A Bayesian Modelling Framework: Concepts, Structure, and Extensibility». *Statistics and Computing*, 10 (4), 325-337. <https://doi.org/10.1023/A:1008929526011>

MATLAB (2019). «Version 9.7 (R2019b)». The MathWorks Inc.

NELDER, J. A. y MEAD, R. (1965). «A simplex method for function minimization». *The Computer Journal*, 7, 308-313. <https://doi.org/10.1093/comjnl/7.4.308>

PICCARRETA, R. (2017). «Joint sequence analysis: Association and clustering». *Sociological Methods & Research*, 46 (2), 252-287. <https://doi.org/10.1177/0049124115591013>

PICCARRETA, R. y MATTHIAS, S. (2019). «Holistic analysis of the life course: Methodological challenges and new perspectives». *Advances in Life Course Research*, 41, 100251. <https://doi.org/10.1016/j.alcr.2018.10.004>

PLUMMER, M. et al. (2003). «JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling». *Proceedings of the 3rd International Workshop on Distributed Statistical Computing*, 124 (125.10), 1-10.

POL, F. de y LANGEHEINE, R. (1990). «Mixed Markov latent class models». *Sociological Methodology*, 213-247. <https://doi.org/10.2307/271087>

PRENTICE, R. L.; KALBFLEISCH, J. D.; PETERSON, A. V.; FLOURNOY, N.; FAREWELL, V. T. y BRESLOW, N. E. (1978). «The analysis of failure times in the presence of competing risks». *Biometrics*, 34 (4), 541-554. <https://doi.org/10.2307/2530374>

R Core Team (2019). *R: A Language and Environment for Statistical Computing*. <https://www.R-project.org/>.

ROSSIGNON, F.; STUDER, M.; GAUTHIER, J. A. y LE GOFF, J. M. (2018). «Sequence history analysis (SHA): Estimating the effect of past trajectories on an upcoming event». En: RITSCHARD, G. y STUDER, M. (eds). *Sequence Analysis and Related Approaches. Life Course Research and Social Policies*, 10. Cham: Springer. <https://doi.org/10.1007/978-3-319-95420-2_6>

RUE, H.; MARTINO, S. y CHOPIN, N. (2009). «Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations». *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 71 (2), 319-392. <https://doi.org/10.1111/j.1467-9868.2008.00700.x>

SANTONJA, F. J.; TARAZONA, A. C., & VILLANUEVA, R. J. (2008). «A mathematical model of the pressure of an extreme ideology on a society». *Computers & Mathematics with Applications*, 56 (3), 836-846. <https://doi.org/10.1016/j.camwa.2008.01.001>

SCHERER, S. (2001). «Early career patterns: A comparison of Great Britain and West Germany». *European Sociological Review*, 17 (2), 119-144. <https://doi.org/10.1093/esr/17.2.119>

SCOTT, S. L. (2002). «Bayesian methods for hidden Markov models: Recursive computing in the 21st century». *Journal of the American statistical Association*, 97(457), 337-351. <https://doi.org/10.1198/016214502753479464>

SIMÓ, C.; BONMATÍ, A. y GOLSCH, K. (2006). «Globalization and men's mid-career occupational mobility in Spain». En: *Globalization, Uncertainty and Men's Careers: An International Comparison*.

STUDER, M.; LIEFBROER, A. C. y MOOYAART, J. (2017). «Understanding Trends in the Transition to Adulthood: An Application of Competing Trajectories Analysis». *PAA 2017 Annual Meeting*.

STUDER, M.; STRUFFOLINO, E. y FASANG, A. E. (2018). «Estimating the relationship between time-varying covariates and trajectories: The sequence analysis multistate model procedure». *Sociological Methodology*, 48 (1), 103-135. <https://doi.org/10.1177/0081175017747122>

STURTZ, S.; LIGGES, U. y GELMAN, A. (2005). «R2WinBUGS: A Package for Running WinBUGS from R». *Journal of Statistical Software*, 12 (3), 1-16. <https://doi.org/10.18637/jss.v012.i03>

SU, Y. S. y YAJIMA, Masanao (2020). *R2jags: Using R to Run "JAGS."* <https://CRAN.R-project.org/package=R2jags>.

THERNEAU, T. (2021). «A Package for Survival Analysis in R». <https://CRAN.R-project.org/package=survival>.

THERNEAU, T.; CROWSON, C. & ATKINSON, E. (2020). *Multi-state models and competing risks. CRAN-R*

VERD, J. M.; BARRANCO, O. y BOLÍBAR, M. (2019). «Youth unemployment and employment trajectories in Spain during the Great Recession: what are the determinants?». *Journal for Labour Market Research*, 53 (1), 4. <https://doi.org/10.1186/s12651-019-0254-3>

VERMUNT, J. K. (2008). «Latent class and finite mixture models for multilevel data sets». *Statistical Methods in Medical Research*, 17 (1), 33-51. <https://doi.org/10.1177/0962280207081238>

VISSER, Ingmar y SPEEKENBRINK, Maarten (2010). «depmixS4: An R Package for Hidden Markov Models». *Journal of Statistical Software*, 36 (7), 1-21. <https://doi.org/10.18637/jss.v036.i07>

WASSERSTEIN, R. L.; SCHIRM, A. L. y LAZAR, N. A. (2019). «Moving to a World Beyond "p < 0.05"». *American Statistician*, 73 (sup1.), 1-19. <https://doi.org/10.1080/00031305.2019.1583913>

WU, L. L. (2000). «Some comments on "Sequence analysis and optimal matching methods in sociology: Review and prospect"». *Sociological Methods & Research*, 29 (1), 41-64. <https://doi.org/10.1177/0049124100029001003>