Guia d'anàlisi longitudinal FUNDACIÓ JAUME BOFILL

Guia introductòria a l'anàlisi longitudinal de dades de panel

Exemples pràctics a partir del Panel de Desigualtats Socials a Catalunya- PaD

Francesc X. Belvis Costes i Joan Benach i Rovira (coordinadors)

NOVEMBRE 2013

Obra Social "la Caixa"

FUNDACIÓ JAUME BOFILL

PaD
Panel de Desigualtats a Catalunya

Aquest treball està subjecte a la llicència
Creative Commons de Reconeixement-
NoComercial-SenseObraDerivada

Les publicacions de la Fundació Jaume Bofill
estan disponibles per a descàrrega al web
www.fbofill.cat.

© Fundació Jaume Bofill, 2014
Provença, 324
08037 Barcelona
fbofill@fbofill.cat
http://www.fbofill.cat

Primera edició: novembre de 2014
Coordinació: Francesc X. Belvis Costes
i Joan Benach i Rovira
Autoria
- Anàlisi estadística: Francesc X. Belvis Costes, Sergio Salas Nicás i Montse Vergara Duarte
- Redacció i revisió: Francesc X. Belvis Costes, Sergio Salas Nicás, Montse Vergara Duarte, Josep Anton Sánchez Espigares i Albert Navarro Giné

Col·laboració: Irene Galí Magallón i resta de l'equip GREDS-EMCONET
Coordinació de continguts: Laura Morató i Vázquez
Cura editorial: Ruth Cortés
Imatge de la coberta: bloc.pucp.edu.pe

Índex

Pròleg

La *Guia introductòria a l'anàlisi longitudinal de dades de panel* que us fem a mans és fruit de diverses realitats de les quals hem anat prenent consciència durant els onze anys de vida del Panel de Desigualtats Socials a Catalunya (PaD).

El PaD va ser concebut com una base de dades longitudinal amb una mostra panel però, contràriament al que esperàvem, el nombre de recerques que han aprofitat el potencial d'aquest tipus de dades i que han fet ús de tècniques d'anàlisi longitudinal ha estat relativament baix.

Pensem que això és degut a diversos motius. D'una banda, hi ha molt poca tradició en estudis longitudinals al nostre país. Aquest tipus d'estudis, que en altres països (com els Estats Units, la Gran Bretanya o Alemanya, per posar-ne alguns exemples) gaudeixen de reconeixement en el món acadèmic i polític des de fa dècades, a casa nostra tenen un pes gairebé imperceptible. D'altra banda els futurs investigadors, en el seu pas per les universitats catalanes, no reben formació específica sobre anàlisi longitudinal. La poca tradició d'aquest tipus d'anàlisis, unida a la poca formació existent, fa que els investigadors interessats a endinsar-se en un món nou, de major complexitat analítica, però també de major potencial explicatiu, es redueixin de manera important. Cal anar a buscar la tradició i la formació a l'estranger.

La Fundació Jaume Bofill, interessada en l'estudi de l'estructura social catalana i, en especial, de les desigualtats socials, ja va constatar aquestes mancances fa més d'una dècada. Per això va impulsar el naixement del PaD, un instrument estadístic per recollir dades longitudinals, alineant Catalunya amb els països més avançats del món pel que fa a la recerca social. Va apostar per una base de dades que permetés l'extensió d'una anàlisi rica, aprofundida i complexa de la realitat social. Era el primer pas.

En aquests moments, amb el treball de camp del PaD en hibernació, volem mirar d'ajudar a aprofitar-ne la vessant de l'anàlisi longitudinal amb l'edició d'aquesta Guia. Ens ha esperonat a fer-ho, també, comptar amb la col□laboració del grup de recerca GREDS-EMCONET de la UPF. L'alt nivell metodològic i tècnic de la seva feina ens ha fet decidir a posar negre sobre blanc una guia per a l'anàlisi longitudinal que pogués ser un punt de partida, un referent inicial, per a l'extensió de l'estudi dinàmic de la realitat en el món acadèmic.

Laura Morató i Vázquez
Responsable de l'àrea de dades del PaD
lmorato@fbofill.cat

Presentació

En el marc de les ciències socials, l'anàlisi transversal de les dades continua sent l'enfocament més àmpliament utilitzat. Fins fa poc això era perfectament coherent amb els plantejaments de recerca i també, correlativament, amb la natura transversal de la gran majoria de bases de dades existents.

I els dissenys longitudinals? En primer lloc, cal insistir que no constitueixen cap panacea des del punt de vista de la investigació, essent més o menys adequats en funció dels objectius i les possibilitats de la recerca. Tanmateix, des de fa molt temps sabem que ofereixen *ceteris paribus* alguns avantatges en relació als transversals, com ara una major evidència a l'hora d'establir l'existència de relacions causals, mentre que resulten imprescindibles quan es tracta d'estudiar el canvi intraindividual.

Al llarg de les dues darreres dècades hem vist un increment notable del nombre de publicacions que reclamen el qualificatiu de "longitudinals" (Taris, 2000, p. VII), mentre augmentava també la disponibilitat de bases de dades aptes per aquesta mena d'anàlisi. A banda del creixent interès teòric pels aspectes dinàmics de la vida social, cal emfasitzar el paper que les noves tecnologies de la informació i la comunicació han jugat en aquesta evolució. L'increment exponencial de la potència de càlcul dels ordinadors ha tornat factibles tècniques estadístiques abans impensables de dur a la pràctica, alhora que ha estimulat la recerca estadística en aquest àmbit. La capacitat d'emmagatzematge d'informació i les facilitats de manipulació de la mateixa -molt particularment, la capacitat de vincular dades procedents de diverses fonts-, obren possibilitats d'autoconeixement de la societat que actualment tan sols comencen a desenvolupar-se. Pensem per exemple com la informatització aplicada als diversos sectors de l'administració pública ha permès crear tant a nivell nacional com internacional reservoris de bases de dades longitudinals sobre aspectes com ara l'educació, la vida laboral o la salut, que obren potencialitats enormes per a la recerca.

Malauradament, la disponibilitat de dades longitudinals s'ha estès més ràpidament que el coneixement de les tècniques estadístiques adequades per a la seva anàlisi, a pesar de que molts avenços comparativament recents de la teoria estadística en aquest camp estan ja implementats en paquets de *software* estadístic d'orientació general. Per exemple, en una revisió de la literatura, Halaby (2004) constatava la lentitud de la sociologia a l'hora d'aprofitar els avantatges que ofereixen les dades de panel per controlar l'efecte de les variables no observades, quelcom que amenaça la validesa de les inferències causals en els estudis de tipus observacional.

En el context de la sociologia i la politologia que es fan a Catalunya i l'Estat Espanyol, encara hi ha un aprofitament escàs de les possibilitats contingudes en les bases de dades longitudinals. Sovint la recerca opta per ignorar la natura longitudinal de la informació, tot limitant-se a dissenys de tipus transversal. Fins i tot quan el plantejament de recerca és pròpiament longitudinal, el desconeixement del ventall de tècniques estadístiques actualment disponibles i les condicions de la seva aplicació, més la seva complexitat intrínseca afegida a l'escassa estandardització de la nomenclatura en un àrea en constant evolució, fan que s'opti sovint per tècniques que no aprofiten tot el potencial de les dades -com ara anàlisis limitades a dos punts en el temps.

Un exemple clar d'aquesta situació és l'escàs aprofitament que des del punt de vista longitudinal ha rebut el Panel de Desigualtats Socials a Catalunya (PaD) de la Fundació

Jaume Bofill, el qual proveeix una matèria primera del màxim interès per a la recerca en diversos àmbits com ara el laboral, social, polític i educatiu. Fins ara les seves dades han estat estudiades fonamentalment des d'un punt de vista transversal, mentre que els dissenys longitudinals s'han limitat a l'aplicació de tècniques d'anàlisi comparativament simples. Evidentment això no invalida les troballes ni conclusions ben establertes, però no dubtem que un coneixement més aprofundit de les possibilitats d'anàlisi actualment a disposició dels investigadors i investigadores hauria conduït a triar altres opcions en relació a l'anàlisi estadística, i qui sap si ha estimular més plantejaments de recerca longitudinal.

Encara que el mercat editorial ofereix bons manuals de tècniques d'anàlisi longitudinal, detectem una certa manca d'adaptació dels mateixos a les necessitats i els plantejaments més habituals en el camp de la sociologia i la politologia. Efectivament, cada branca del coneixement científic (econometria, psicologia, medicina, sociologia...) sol construir el seu propi equipament adaptat d'eines estadístiques, i això resulta també cert pel que fa a les tècniques d'anàlisi longitudinal (Rabe-Hesketh i Skrondal, 2012a, p. 229). En el marc de les ciències socials, l'econometria disposa de textos excel·lents per a l'anàlisi de les dades de panel però escassament adaptats al context de la sociologia i la politologia. Això es deu tant al seu alt nivell matemàtic com al fet que les tècniques seleccionades estan exclusivament orientades a la variable i posen l'èmfasi en l'obtenció d'evidència causal.

Aquesta Guia introductòria vol contribuir modestament a popularitzar la recerca longitudinal, tot familiaritzant els investigadors i investigadores amb les tècniques d'anàlisi que es troben al seu abast en aquest àmbit. Tres són les seves aportacions principals. En primer lloc, proposem una classificació de les tècniques d'anàlisi de dades longitudinals junt amb algunes referències bibliogràfiques que ajudaran a orientar-se quant a les seves característiques, adequació, fortaleses i limitacions. D'aquesta manera, qualsevol investigador podrà triar d'una forma raonada les eines més adients per fer front a les necessitats de la seva recerca. En segon lloc, oferim indicacions sobre aspectes generals de l'anàlisi longitudinal, com ara la manera d'organitzar les dades, els problemes relacionats amb la no-resposta o les possibles anàlisis de tipus descriptiu. Finalment, s'ofereixen tres exemples pràctics d'anàlisi longitudinal aplicats a problemes de tipus social, tot utilitzant tècniques seleccionades i diferents programes d'anàlisi estadística de tipus generalista (R, Stata), que poden servir de model i motivació per tal que els investigadors utilitzin amb més freqüència dissenys longitudinals en les seves recerques.

Òbviament aquesta Guia no substitueix els molts i excel·lents manuals disponibles de caire general sobre l'anàlisi longitudinal, ni tampoc aquells que tracten tècniques d'anàlisi específiques. La seva finalitat és introduir a professionals sèniors sense experiència i a recercadors joves interessats en el tema, obviant moltes de les seves complexitats matemàtiques. A pesar d'això, és recomanable que el lector o lectora estigui prèviament familiaritzat amb els conceptes bàsics de l'anàlisi estadística de dades i la construcció de models.

Objectius

L'objectiu general d'aquesta Guia és presentar d'una forma pràctica i tan entenedora com sigui possible les tècniques i els mètodes estadístics principals per a l'anàlisi longitudinal de dades, des de la perspectiva de les ciències socials i en relació als estudis de panel. Per tant, l'objectiu no és aprofundir en els mètodes, sinó introduir-vos-hi i, mitjançant exemples, mostrar-ne les potencials aplicacions. Amb aquesta finalitat utilitzem dades del PaD de la Fundació Jaume Bofill, un panel força conegut pels investigadors, fàcilment accessible i les potencialitats del qual des d'un punt de vista longitudinal han estat escassament explotades.

Els objectius específics de la Guia són quatre:

Caracteritzar el ventall de tècniques disponibles a l'hora d'analitzar dades des d'un punt de vista longitudinal.

Orientar sobre els problemes comuns dels estudis longitudinals, com ara l'organització de les dades, el seu caràcter balancejat o no, o els casos perduts (*missing data*).

Oferir exemples detallats d'aplicació d'aquestes tècniques amb diferents tipus de programari.

Realitzar les anàlisis sobre dades procedents d'un estudi de panel referit al nostre context social pròxim, com és el cas del PaD.

Com s'organitza la Guia

La Guia s'estructura en quatre capítols principals. En primer lloc, un capítol introductori defineix el que entenem per dades longitudinals, els seus avantages i inconvenients, i caracteritza els estudis de panel.

En el segon capítol es descriuen les principals consideracions estadístiques per poder analitzar dades de tipus panel. D'una banda, es mostren aspectes comuns de l'anàlisi, com l'estructura de les dades o la importància de la pèrdua d'informació, molt habitual en els estudis longitudinals. De l'altra, en un segon apartat s'enumeren i es descriuen de manera general les tècniques estadístiques més utilitzades en l'àmbit de les ciències socials per analitzar aquest tipus de dades.

En el tercer capítol es presenten mitjançant estudis de cas algunes de les tècniques descrites en el segon capítol. Aquests exemples inclouen tres casos pràctics aplicats amb diferents conjunts de dades seleccionats del PaD. Les anàlisis s'han realitzat utilitzant dos paquets d'orientació generalista i força difosos com són Stata i l'R (aquest darrer de programari lliure). Cadascun d'aquests exemples inclou de forma breu la conceptualització i els objectius de l'estudi, la descripció de les variables i dades d'estudi, la sintaxi o els comandaments utilitzats en cadascun dels procediments, així com una breu descripció i interpretació de les sortides i els resultats que ofereixen aquests comandaments.

Finalment, la Guia inclou també les referències bibliogràfiques citades al llarg del text, que considerem alhora una selecció bibliogràfica útil per aprofundir en el coneixement de les tècniques d'anàlisi longitudinal.

1. Introducció a les dades longitudinals

Com el seu títol indica, aquesta Guia està focalitzada en l'anàlisi de dades longitudinals. Naturalment, l'existència de dades longitudinals implica l'existència prèvia d'un plantejament de recerca longitudinal, d'un disseny longitudinal i de la seva posada en pràctica en el treball de camp. Tanmateix, els aspectes previs a l'anàlisi de les dades no són l'objectiu d'aquesta Guia i només ens hi referirem tangencialment, en tant que siguin necessaris per a entendre certs aspectes de les dades.

1.1. Dades longitudinals i dissenys longitudinals

Entenem per dades longitudinals aquelles que han estat construïdes i organitzades de manera que donen informació: 1) sobre una o més variables, 2) en almenys dos[1] moments diferents i identificats del "temps", i 3) mesurades sobre la mateixa unitat o unitats d'anàlisi, també identificades.

Es pot parlar indistintament de dades longitudinals o dades de panel, encara que la primera expressió és més freqüent en bioestadística i la segona en economia.[2] Les unitats d'anàlisi habitualment seran persones, encara que pot tractar-se de qualsevol altra entitat, com ara llars, empreses o nacions. En aquest text parlarem indistintament d'unitats d'anàlisi o individus, en un sentit general.

Les dades longitudinals s'oposen sovint a les dades transversals, que es diferencien de les primeres per donar informació sobre les unitats d'anàlisi en un únic moment del temps (Taris, 2000, p. 1).

La definició anterior també exclou les dades transversals repetides (*repeated cross sectional* o *trend data*), en tant que les unitats d'anàlisi són diferents en cada ocasió de mesura. Tanmateix unes dades transversals repetides al nivell d'una certa unitat d'anàlisi (per exemple persones) es poden construir com a longitudinals si canviem d'unitat d'anàlisi (per exemple agregant per centre, regió o país) (Taris, 2000, p. 2).

Quan s'estudia una única unitat d'anàlisi, les dades longitudinals es coneixen com a sèries temporals.

Quan es combinen diverses sèries temporals corresponents a diverses unitats d'anàlisi (*pooled time-series data*), obtenim una estructura de dades formalment indistingible de qualsevol conjunt de dades longitudinals corresponents a més d'una unitat d'anàlisi. Tanmateix, en la pràctica aquestes estructures presenten típicament menys casos i més observacions que la resta de dades longitudinals (Finkel, 1995, p. 90), cosa que condiciona les tècniques estadístiques a fer servir (Baum, 2006, p. 236-242; Menard, 2002, p. 31).

Per aquesta raó se sol distingir, dintre de les estructures de dades longitudinals amb

[1] Les dades longitudinals que tan sols mesuren dos moments presenten diverses limitacions en l'estudi, cosa que porta a Singer i Willett (Singer i Willett, 2003, p. 10) a restringir el qualificatiu de *dades longitudinals* a aquelles que mesuren almenys tres ocasions. Aquí hem considerat més coherent incloure-hi també les de dues mesures.
[2] I d'altra banda l'expressió *dades de panel* connota el disseny de panel, que és només un d'entre els dissenys capaços de generar dades de panel i que els dóna certes particularitats.

múltiples individus, entre panels verticals (*short panel*) i panels horitzontals (*long panel*). Els primers tenen un nombre de casos gran i un nombre d'ocasions de mesura comparativament petit; mentre que els segons presenten un nombre de casos petit i un nombre d'ocasions de mesura comparativament gran (Cameron i Trivedi, 2010, p. 230).

Les estructures de dades longitudinals poden diferir entre si en molts altres aspectes pertinents per a l'anàlisi. Les fonts d'aquestes diferències són bàsicament: *a)* el disseny de recerca que ha generat les dades, i *b)* les realitats del treball de camp.

Les dades longitudinals es produeixen en el si de diversos dissenys de recerca. Aquests divergeixen en criteris diversos, com ara el seu caràcter experimental o observacional, el nombre de mesures previstes i la distribució de les mateixes, el seu caràcter prospectiu o retrospectiu, o bé el grau d'heterogeneïtat de la població estudiada. Cadascun d'aquests dissenys imposa determinades limitacions a les dades longitudinals que podem construir i, per tant, a les anàlisis estadístiques possibles, tal i com veurem més endavant.

1.2. Els dissenys de panel

Per estudis de panel s'entén generalment un tipus de disseny longitudinal observacional, on els individus d'una mostra transversal d'una població són entrevistats o mesurats a intervals generalment regulars (per exemple, cada sis mesos o cada any) i durant un període "llarg" de temps.
Aquest és un disseny àmpliament utilitzat en l'àrea de les ciències socials. Alguns exemples internacionals destacables són el Panel Study of Income Dynamics (PSID) als Estats Units, el British Household Panel Survey (BHPS) al Regne Unit, el Household, Income and Labor Dynamics in Australia (HILDA) a Austràlia, el Socio-Economic Panel Study (GSOEP) a Alemanya o el Swiss Household Panel (SHP) a Suïssa.

Tots aquests panels tenen com a objectiu principal estudiar les dinàmiques generals socials, demogràfiques i econòmiques en els seus països respectius. En aquest context, és molt freqüent que la unitat de mostreig primària siguin les llars i no pas els individus, per la qual cosa se sol utilitzar l'expressió panel de llars (*household panel*) per referir-s'hi. Altres panels se centren en grups específics de la població, per exemple infants, joves o treballadors (Lee, 2003).
A Espanya existeix un dèficit molt notable d'estudis de panel. L'exemple més conspicu és l'Encuesta de Población Activa (EPA), tot i que es tracta d'un panel "rotatori" (*revolving panel design*) (Menard, 2002, p. 28), en què els individus són seguits longitudinalment només durant un període de temps curt, un any i mig com a màxim (Jimenez-Martin i Peracchi, 2002).

A Catalunya, la Fundació Jaume Bofill ha dut a terme entre el 2002 i el 2012 el Panel de Desigualtats Socials, un panel de llars que és la font de les dades que utilitzem en els exemples pràctics d'aquesta Guia. A l'inici del capítol 3 hom pot trobar una caracterització més exhaustiva del PaD.

Típicament, els estudis de panel generen estructures de dades de tipus panel vertical (short *panel*), amb les unitats d'anàlisi mesurades a intervals constants i (almenys, en el context de les ciències socials) una proporció significativa de valors perduts. El PaD no és una excepció en aquest sentit. Encara que els exemples pràctics d'aquesta Guia es

basen en dades generades a partir d'un únic disseny, tant les tècniques emprades com les explicacions de caire més general que es proporcionen són extrapolables amb pocs o cap canvi a dades procedents d'altres tipus de dissenys longitudinal.

1.3. Inconvenients i potencial de les dades longitudinals

Els inconvenients de les dades longitudinals rauen fonamentalment en les qüestions 1) pràctiques, i 2) de qualitat de les dades.

1) Des del punt de vista pràctic, l'obtenció de dades longitudinals és sovint més complicada, i per tant més costosa, que la de dades transversals. En l'àmbit de les ciències socials això és tradueix en dissenys de recerca que requereixen un gran pressupost, només a l'abast d'unes poques institucions. L'investigador o investigadora es veu normalment reduït a l'anàlisi secundària i a un ventall limitat de temes o/i possibilitats d'operacionalitzar-ne els conceptes.

Cal apuntar tanmateix que la tendència actual a la informatització i l'enllaçament dels registres dels actes administratius més diversos, particularment en institucions públiques dels àmbits de l'educació, la sanitat, la cultura, el treball, etc. (i en la mesura que són posades a disposició dels investigadors), constitueixen una font emergent extremadament rica de dades longitudinals d'una qualitat elevada i sobre problemàtiques d'abast mitjà que conviden particularment a l'aplicació de tècniques longitudinals.

2) Des del punt de vista de la qualitat de les dades, els dissenys longitudinals estan sotmesos als mateixos problemes de validesa i fiabilitat que afecten els dissenys transversals, als quals cal afegir-n'hi alguns d'específics. En el cas dels dissenys prospectius, s'ha constatat l'existència d'un efecte de reactivitat (*panel conditioning*), provocat per l'aplicació repetida del mateix instrument de mesura als mateixos subjectes; alternativament, en els dissenys retrospectius apareixen biaixos lligats a la memòria dels informants (Menard, 2002, p. 38-42).

Les reflexions sobre la validesa podrien portar-nos també a qüestionar en algunes ocasions la comparabilitat de les mesures al llarg del temps. Els fets socials són dinàmics a molt més curt termini que altres aspectes de la realitat i, en aquest sentit, no sempre podem garantir que les variables mesurades representin dimensions estables de la realitat: sota aquest supòsit, la comparació de mesures d'un mateix individu en el temps perd tot sentit, o n'adquireix un de nou. Per exemple, les implicacions de posseir un contracte fix poden ser molt diferents abans i després d'un canvi legislatiu, de manera que, encara que sigui nominalment idèntica, la categoria no representa el mateix. Aquestes qüestions són molt rellevants però queden fora de l'abast d'aquesta Guia, per la qual cosa remetem únicament a una presentació del tema (Taris, 2000, p. 39-44).

Els valors perduts, en canvi, són un aspecte de la qualitat de les dades longitudinals extremadament pertinent des del punt de vista estadístic. L'anàlisi longitudinal de dades és sensible en major grau als efectes de la no-resposta, i aquesta és una questió que apareixerà repetidament en aquesta Guia.

L'anàlisi longitudinal no és, per tant, cap panacea, i ni tan sols és imprescindible per a la majoria de qüestions de recerca. Però, malgrat les consideracions anteriors, la dispo-

nibilitat de dades longitudinals té dos avantatges clars en relació amb les transversals:

1) De forma general, poden proveir un major grau d'evidència respecte a l'existència d'efectes causals (Finkel, 1995, p. 1).

Efectivament, en el context de l'anàlisi de la covariació (i.e. dades procedents de dissenys no experimentals), el gran enemic de la interpretació causal de les associacions descobertes en un model de regressió és l'existència d'endogeneïtat o correlació entre les variables predictores i el terme d'error de la regressió. En presència d'endogeneïtat, ni els coeficients estimats són consistents, ni els seus errors estàndard precisos.

L'endogeneïtat pot tenir diverses fonts: *a)* heterogeneïtat individual, *b)* error de mesura en les variables independents, i *c)* determinació recíproca entre la variable dependent i les independents. En tots aquests casos, les dades longitudinals contenen sovint informació suficient per afrontar aquests problemes, una possibilitat que no es presenta en el cas de les dades transversals.

- Quant a l'heterogeneïtat individual, en un context experimental l'assignació aleatòria dels individus al grup experimental o de control torna els grups equivalents, però amb dades observacionals és evident que els individus poden diferir entre si en moltes característiques, a banda de les controlades en el model de regressió (Allison, 2005, p. 1). Si aquestes característiques no observades estan relacionades amb les variables explicatives, aquestes seran endògenes.

La disponibilitat de dades longitudinals permet comparar cada individu amb si mateix. D'aquesta manera cada individu actua com el seu propi control, i l'efecte de les característiques específiques del subjecte queda eliminat (Rabe-Hesketh i Skrondal, 2012a, p. 244).

- De forma similar, l'existència de mesures repetides per a un mateix individu obre la possibilitat de controlar diverses formes de l'error de mesura, com ara el fenomen de la *regressió a la mitjana* (Finkel, 1995, p. 8-9, 47-48).

- Finalment, molts processos socials es caracteritzen per la causació recíproca de les variables analitzades. Atès que la causa ha de precedir necessàriament l'efecte, la dimensió temporal implícita en les dades longitudinals permet estudiar, per exemple mitjançant la introducció de valors previs de les variables independents o dependent, camins causals que són generalment indecidibles en el cas de les dades transversals. La coneguda distinció entre els efectes d'edat, període i cohort seria un exemple particularment rellevant d'aquesta propietat de les dades longitudinals.

2) Si l'objectiu del nostre estudi és el canvi intraindividual, la disponibilitat de dades longitudinals és pràcticament imprescindible. Aquestes fan possible l'estudi de les trajectòries individuals, la predicció de les mateixes i (potser més interessant des del punt de vista de la ciència social) l'existència de diferències interindividuals en les trajectòries intraindividuals.

2. L'anàlisi estadística longitudinal de les dades de panel

2.1. Aspectes comuns

2.1.1. L'organització de les dades: format "llarg" vs. format "ample"

A l'hora d'organitzar una taula de dades de tipus longitudinal, existeixen bàsicament dos formats: el format "llarg", i el format "ample". El format llarg (*long, person-period dataset*) és aquell en què cada columna de la taula de dades correspon a una variable diferent, per exemple, l'edat, el sexe, el nivell d'estudis, o l'esdeveniment d'interès que volem estudiar, com ara els ingressos, el tipus de contracte laboral, etc.

Atès que cada individu ha estat mesurat repetidament en aquestes variables, existiran múltiples registres ("files") per individu –un per cada ocasió de mesura, i es fa necessària una variable addicional identificadora del temps (o número d'ocasió) en què l'individu ha estat mesurat.

D'altra banda, el format ample (*wide, person-level dataset*) és aquell en què cada columna representa les diferents ocasions de mesura de la mateixa variable. Per exemple, si s'han pres deu mesures sobre el pes d'una persona durant un període de seguiment, hi hauria deu variables diferents de pes, corresponents a cadascuna de les mesures. En aquest cas, només existirà un únic registre ("fila") per persona.

Indubtablement, el format llarg resulta més flexible i és la forma més corrent d'estructurar les dades en el cas de les anàlisis longitudinals. El format ample és més pràctic per a determinats usos (per exemple, permet calcular directament la correlació entre mesures corresponents a diferents moments del temps), però té dificultats quan els moments d'observació no són els mateixos per a tots els individus, o l'espaiament entre les observacions no és constant. A més, el format llarg permet construir una base de dades més parsimoniosa, en el sentit que el nombre de columnes de la mateixa es redueix considerablement.

En ocasions ens podem trobar estructures de dades de tipus llarg més complexes que l'estructura simple descrita aquí. Estem pensant particularment en el context de l'anàlisi de supervivència, quan es volen analitzar esdeveniments repetits per subjecte, o bé múltiples esdeveniments (Javier i Bart, s.d.); però el principi d'organització de les dades es manté.

Els paquets estadístics acostumen a tenir implementada l'opció de traspàs de format llarg a format ample.

2.1.2. La no- resposta en els dissenys longitudinals

Per no-resposta entenem la no-obtenció d'un valor informatiu[3] en la mesura d'un ítem programada en el disseny de la recerca (Fitzmaurice, Davidian, Verbeke i Molenberghs,

[3] Si el valor en qüestió es considera informatiu o no, és una decisió de l'investigador o investigadora. Per exemple, l'opció "No sap" o "No contesta" en relació amb el vot en les darreres eleccions pot ser un valor vàlid en un determinat plantejament de recerca, mentre que constitueix una manca d'informació en un altre.

2009, p. 395).

La no-resposta implica sempre una reducció en la grandària efectiva de la mostra, per a algunes o per a totes les anàlisis previstes en el disseny. Si a més la reducció no s'ha produït segons un patró completament aleatori en relació amb les variables d'interès, la no-resposta ocasionarà també pèrdua de representativitat de la mostra.

En conseqüència, la no-resposta fa que les estimacions obtingudes únicament a partir dels casos informats tinguin una precisió menor de la inicialment prevista en el disseny. Si a més ha ocasionat pèrdua de representativitat, la no-resposta induirà també un biaix en les estimacions obtingudes. Aquesta segona possibilitat és habitualment la més inquietant.

En relació amb la no-resposta, és útil distingir entre dues casuístiques:

No-resposta d'ítem (*item non-response*). La no-resposta afecta un subconjunt limitat d'ítems.

No-resposta d'unitat d'anàlisi (*unit non-response*). La no-resposta afecta tots els ítems del disseny. En el cas dels estudis longitudinals, la no-resposta d'unitat d'anàlisi pot ser definitiva o temporal.

Conceptualment, la diferència entre ambdues casuístiques és només quantitativa. La primera situació s'apropa a la segona a mesura que, per a una determinada unitat d'anàlisi, el nombre d'ítems per als quals manca informació s'incrementa.[4] En la pràctica, tanmateix, sol existir un tall qualitatiu evident entre una situació i l'altra.

La no-resposta d'ítem i d'unitat afecta tant els dissenys transversals com longitudinals. En aquests darrers, però, revesteix una major gravetat, simplement perquè el seu desplegament al llarg del temps multiplica les ocasions que s'originin valors perduts i aquests adquireixen un caràcter acumulatiu (Fitzmaurice *et al.*, 2009, p. 395).

2.1.2.1. El desgast mostral (*attrition*)

En els dissenys transversals la no-resposta d'unitat d'anàlisi equival a la no-participació de l'individu un cop ha estat seleccionat per formar part de la mostra. Aquesta és una qüestió a la qual en general es dóna poca rellevància, especialment si tenim en compte les taxes elevades de refús, que són habitualment d'entre el 30% i el 40%. Els efectes d'aquest procés de selecció s'intenten compensar mitjançant la substitució d'individus, i (*a posteriori*) mitjançant la ponderació de la mostra, però la primera mesura no garanteix l'equivalència dels individus i la segona només permet controlar el biaix en determinades variables.

En el cas dels dissenys longitudinals, els efectes de la no-resposta d'unitat d'anàlisi resulten més evidents i potencialment molt greus, a causa del seu caràcter acumulatiu. El desgast mostral (*attrition*) és un fenomen propi dels dissenys longitudinals que es defi-

[4] Per exemple en les enquestes d'estudis de panel a llars (com és el cas del PaD), és habitual recollir la informació referida a membres de la família absents mitjançant un questionari proxy, que és respost per l'informant principal de la llar i que consisteix en una selecció dels ítems programats en el questionari inicial. L'emplenament del questionari proxy equival per tant a la no-resposta sistemàtica en un nombre considerable d'ítems.

neix com la impossibilitat d'entrevistar o mesurar una unitat mostral prèviament seleccionada, en qualsevol de les rondes programades al llarg del període d'observació (Peracchi, 2002).

L'*attrition* genera diferents perfils d'individus: si la pèrdua és definitiva parlem d'individus *dropouts* i en el cas contrari parlem de *returners*. Els mesurats en totes les ocasions programades són *completers*.

Des del punt de vista quantitatiu, l'*attrition* és la principal font de valors perduts. S'ha realitzat una infinitat d'estudis que analitzen els determinants de l'*attrition*, sobretot al nivell internacional (Behr, Bellgardt i Rendtel, 2005; Lee, 2003). Aquests determinants poden variar en funció de l'objectiu de l'estudi o el seu context, és a dir, el país on es realitza, així com la població d'estudi. Alhora, els determinants poden ser diferents en funció de si s'analitza la pèrdua d'individus o bé de tota la llar. En aquesta Guia s'inclou un cas pràctic d'estudi del desgast mostral, aplicat a identificar els factors determinants de la pèrdua de llars.

La probabilitat d'*attrition* depèn freqüentment de variables molt influents des del punt de vista social. Per exemple, Van Berg, Lindeboom i Ridder (1994) troben que el comportament dels individus respecte al mercat de treball i l'*attrition* del panel no són processos independents. És evident que aquestes associacions poden induir biaixos notables en els resultats de qualsevol tècnica estadística, si no s'intenten corregir d'alguna manera.

2.1.2.2. Tractament de la no-resposta

El tractament de la no-resposta o *missing* està rebent una atenció progressiva en la recerca científica. També és una qüestió complexa i sobre la qual existeix nombrosa literatura. En aquesta Guia ens limitarem a descriure alguns dels conceptes essencials que cal tenir presents a l'hora de tractar els casos perduts, així com a oferir algunes referències específiques.

En un treball ja clàssic, Rubin distingí tres mecanismes de creació de valors perduts: valors perduts completament a l'atzar (*missing completely at random*, MCAR), valors perduts a l'atzar (*missings at random*, MAR) i valors perduts no a l'atzar (*missing not at random*, MNAR).

En el mecanisme de MCAR, la pèrdua d'informació en una variable es produeix de manera independent de qualsevol altra variable, ja sigui observada o no observada. Sota aquest supòsit, es produirà una pèrdua de potència, però les anàlisis estadístiques no estaran esbiaixades. Un exemple d'aquest mecanisme seria la no-resposta accidental a una pregunta d'un qüestionari.

En el mecanisme de MAR, la pèrdua d'informació en una variable està associada amb altres (una o més) variables observades del disseny, però un cop condicionada a aquestes, la pèrdua d'informació és MCAR. Aquest seria el cas, per exemple, d'un individu que no es trobava a la llar en el moment de fer l'entrevista, perquè estava treballant.

Finalment, en el mecanisme de MNAR, la pèrdua d'informació es produeix per una raó específica relacionada amb la mateixa variable que presenta valors perduts. Un exemple típic seria la no-resposta en la variable ingressos, si (per exemple) són les persones amb

rendes més altes les que tendeixen a no respondre aquesta qüestió. Aquest mecanisme és el que presenta més complicacions, ja que porta més fàcilment a informacions esbiaixades i no és fàcil trobar els models apropiats per resoldre-ho.

Atès que és inevitable comptar amb alguna pèrdua d'informació als estudis longitudinals, cal tenir presents les diferents estratègies que s'han desenvolupat per reduir-ne les conseqüències en la mesura del possible.

Els procediments tradicionalment utilitzats no es basaven en cap teoria estadística (Collins, 2006). Restringir l'anàlisi als individus amb informació completa (*casewise deletion)* ha estat i continua sent segurament l'estratègia més utilitzada. En realitat, no té tan males propietats, però pot suposar una pèrdua important de potència i, en el cas dels estudis longitudinals, l'acumulació de casos perduts molt probablement crearà l'anomenat "biaix del supervivent", si els *completers* tenen característiques diferents dels *dropouts* en les variables d'interès (Baum, 2006).

Altres estratègies tradicionals, com la imputació del valor mitjà (*mean substitution)*, obtenen uns resultats encara pitjors. En el cas dels estudis longitudinals, la substitució d'una mesura que falta pel valor anterior de l'individu (Last Observation Carried Fordward) o posterior (Next Observation Carried Backward) funciona millor, a causa de la dependència existent entre les mesures individuals. Exhaustives descripcions d'altres estratègies tradicionals de tractament del problema dels valors perduts poden trobar-se a la bibliografia (Allison, 2001; Enders, 2010; Rivero Rodríguez, 2011).

En el present, dues tècniques són considerades *state of the art* quant al tractament del problema dels valors perduts: 1) l'estimació mitjançant la funció de versemblança; i 2) la imputació múltiple (Enders, 2010). Tot i que són més complexes que els mètodes tradicionals, són més eficients i estan basades en conceptualitzacions estadístiques.

Particularment, els mètodes de modelització "moderns" orientats a la variable descrits en aquesta Guia realitzen les estimacions pel mètode de màxima versemblança. Això permet realitzar estimacions consistents dels coeficients de regressió, sempre que el mecanisme de pèrdua sigui MAR (o, naturalment, MCAR). Sota aquest supòsit, aquests mètodes permeten utilitzar tota la informació disponible dels subjectes, encara que presentin un perfil de *dropouts*. Això contrasta amb tècniques tradicionals com el MANOVA, en què els subjectes amb algun valor perdut són exclosos de l'estimació (Rabe-Hesketh i Skrondal, 2012a, p. 278).

2.1.3. Dades equilibrades i no equilibrades

L'"equilibri de les dades" fa referència a la igualtat dels individus quant al *nombre* de les seves observacions, la *simultaneïtat* de les mateixes en el temps i l'*espaiament* (constant o no) entre elles.

D'entrada, aquestes característiques es programen en el disseny de l'estudi, però és la realitat del treball de camp i la recollida de dades la que acaba de configurar-les.

Generalment, al nivell del disseny, el nombre d'observacions es programa com: *a)* específic del subjecte, o bé *b)* un nombre d'observacions fix i comú als individus; en aquest cas és també molt freqüent que la mesura es programi de forma "simultània" per a tots

els subjectes.

En a), l'individu és mesurat cada cop que es produeix un canvi en la variable dependent o les covariables. Hi intervingui o no la no-resposta, aquest plantejament tindrà com a resultat un nombre variable d'observacions per individu. A més, l'espai de temps entre les mesures serà també variable. Aquesta sol ser la situació habitual en els estudis de cohort.

En b), el nombre d'observacions per individu hauria de ser el mateix, com també l'espaiament entre elles. Tanmateix, en la pràctica, l'*attrition* ocasionarà que el nombre de mesures reals per individu variï (gairebé sempre serà així en el marc de les ciències socials). Quelcom similar ocorre amb l'espaiament de les mesures: encara que estigui programat com a constant, l'existència de *returners* o bé el *timing* de les entrevistes fa que l'espaiament variï en la pràctica. Aquesta sol ser la situació habitual en estudis de panel com ara el PaD.

Quan el nombre de mesures és estrictament el mateix per a tots els individus, parlem de dades *fortament equilibrades*. Quan és així per a un gran nombre d'individus, parlem de forma laxa de dades equilibrades. En cas que el nombre de mesures presenti una variabilitat important entre individus parlarem, també de forma laxa, de dades no equilibrades.

El nombre i l'espaiament de les mesures és un dels factors que condiciona el tipus d'anàlisi que podem aplicar a les dades. Algunes tècniques estadístiques requereixen dades estrictament equilibrades. És el cas, per exemple, de l'anàlisi de variança multivariat (MANOVA), o l'anàlisi de seqüències. Això no les descarta per a ser utilitzades en el cas contrari, però certament representa una dificultat que cal tractar. En el cas de reduir l'anàlisi als casos complets, per exemple, el biaix de selecció pot ser molt important.

Altres tècniques no requereixen dades equilibrades sinó que són capaces d'aprofitar la informació aportada per individus *dropouts* en aquelles ocasions en què han estat mesurats, sempre que ni la variable dependent ni cap de les covariables no sigui un valor perdut. Això les fa menys susceptibles al biaix de selecció (Rabe-Hesketh i Skrondal, 2012a, p. 278).

La unitat mínima de mesura del temps

Estrictament parlant, cap fenomen es pot mesurar en temps continu. Alguns fenòmens adopten intrínsecament una forma discreta, com ara el nombre d'entrevista en un estudi panel o bé el nombre de mamografia de la dona en un programa de cribratge. En aquests casos, la unitat mínima ve donada naturalment.

En cas contrari, la unitat mínima de mesura del temps en el nostre estudi hauria de ser, idealment, la més petita possible rellevant en relació amb el fenomen d'interès (Singer i Willett, 2003, p. 313). Per exemple, en relació amb la pèrdua de la feina, una freqüència diària de mesura resultarà segurament excessivament precisa, mentre que una mesura realitzada anualment resultarà massa imprecisa.

En ocasions, la unitat mínima s'arrodoneix per tal d'aconseguir simultaneïtat i/o constància de l'espaiament entre mesures, com ocorre amb els panels, en què rarament

s'utilitza la data d'entrevista, sinó de forma més grollera el nombre d'onada.

De vegades aquest arrodoniment no es tria sinó que ve imposat pel disseny, per la realitat de la recollida de dades o per la impossibilitat de precisar més l'ocurrència dels fenòmens d'interès. Això provoca que, del fenomen d'interès, només se'n conegui què va ocórrer dintre d'un determinat interval; o bé que els possibles canvis ocorreguts en l'interval entre mesures ens siguin desconeguts. En aquestes situacions, es diu que les dades estan censurades a intervals (*interval-censored*).

La unitat mínima de mesura del temps, en determinar la simultaneïtat i/o constància de l'espaiament entre mesures, condiciona també el tipus de tècnica estadística que podem aplicar a les dades. Això és especialment cert en anàlisi de supervivència, en què una mesura comparativament imprecisa del temps fa que sorgeixin molts "empats" entre individus i fa aconsellable, per tant, l'ús de tècniques adaptades al temps discret (Singer i Willett, 2003, p. 315).

2.1.4. L'escala del temps

Les dades longitudinals permeten efectivament seguir els canvis de l'individu en el "temps", però cal advertir que el significat d'aquest terme és ambigu.

En realitat, una decisió crucial en el disseny d'una recerca longitudinal és determinar la naturalesa i el nombre (en cas que en fem servir més d'una) de les escales de temps en la nostra recerca.

En un estudi sobre el desenvolupament de la intel□ligència en infants, sembla natural interpretar l'evolució del coeficient intel□lectual com una funció del pas del temps *interpretat com a increment de l'edat cronològica*. Tanmateix, interpretar l'evolució del salari en una mostra de la població general, per a un determinat període, com una funció de l'edat cronològica resulta bastant més discutible.

En l'estudi sobre l'*attrition* del PaD en aquesta mateixa Guia, el temps està mesurat a temps discret i compta el pas de les successives onades del panel entre el 2002 i el 2009. En un estudi sobre falsos positius en el cribratge del càncer de mama, el temps ve determinat també de forma discreta pel nombre de mamografies que es van realitzant successivament les dones.

Una breu però aclaridora discussió d'aquest tema, juntament amb molts exemples, es pot trobar a Singer i Willett (2003, p. 10-12).

En el marc de les enquestes socials a individus, hi ha generalment tres escales candidates a la mesura del temps: l'edat, la cohort i el període. L'edat representa els possibles efectes de l'etapa del cicle vital en què es troba l'individu; la cohort representa els efectes dels esdeveniments històrics compartits pels membres de la mateixa cohort, mentre que finalment el període representa els efectes dels esdeveniments de l'època en què té lloc la mesura. Malauradament, cadascuna d'aquestes tres mètriques és funció de les altres dues. Edat, període i cohort són perfectament colineals i això impedeix utilitzar-les simultàniament en un model de regressió (Rabe-Hesketh i Skrondal, 2012a, p. 239-241).

En un estudi transversal, el període és fix i no se'n poden estimar els efectes, mentre que els efectes de l'edat i de la generació no poden destriar-se: qualsevol efecte significatiu del "temps" pot interpretar-se alternativament com un efecte de l'edat o de la generació.[5] No podrem distingir, per exemple, si una major religiositat entre la gent gran es deu a l'etapa del cicle vital en què es troben (edat), o bé a l'època històrica en què van néixer (cohort).

En un estudi longitudinal d'una única cohort, els efectes d'aquesta naturalment no es poden estimar, i tampoc no podem distingir entre els efectes de l'edat i el període. Per exemple, podem interpretar un increment dels salaris com a efecte d'un increment de l'experiència (és a dir, un efecte de l'edat laboral) o bé com un efecte del període (a causa de l'actualització dels salaris per la inflació –o fins i tot un increment real dels salaris).

En un estudi longitudinal amb diverses cohorts (*accelerated longitudinal design*), podem estimar els efectes separats de dues mètriques diferents del temps, però aquestes estimacions estan barrejades amb els efectes de la tercera. Només podem interpretar unívocament els coeficients obtinguts si suposem que la tercera mètrica descartada no té efectes significatius sobre la variable dependent. Per exemple, en el segon exemple d'aquesta Guia suposem que els salaris són funció de l'edat i l'experiència, mentre que la cohort no té un efecte significatiu.

2.1.5. Variables "constants" i variables "depenents del temps"

En un estudi longitudinal, la variable dependent necessàriament ha de variar amb el temps a nivell intraindividual –si més no, per a una part dels individus.

Quant a les variables independents, algunes varien intraindividualment amb el temps (*time-varying variables*), mentre que d'altres romanen fixos al llarg del període de seguiment (*time-constant variables* o també *subject-specific variables*).

Aquesta distinció és important en els models longitudinals perquè les variables constants en el temps no poden utilitzar-se en la modelització per explicar la variació intraindividual. Aquesta és la raó per la qual els models d'efectes fixos, centrats exclusivament en l'explicació de la variació intraindividual, no ofereixen estimacions dels efectes de covariables constants.

A l'hora d'analitzar dades longitudinals, caldrà distingir entre aquests dos tipus diferents de variables, i gairebé tots els models que descriurem a la Guia permeten tenir en compte aquesta consideració.

2.1.6. La descripció de les dades

Quan es tracta de dades de panel, les tècniques d'anàlisi descriptiva ordinàries, com ara mesures de resum, forma i dispersió, anàlisis bivariades i multivariants, taules de freqüències, gràfics de tall i fulles (*steam and leaf*) o de caixa (*boxplots)*, etc. aplicades de forma transversal a cada una de les rondes, continuen sent útils per obtenir una primera impressió de la distribució de les variables en estudi, l'evolució de les mateixes en el

[5] Llevat que disposem de l'ajut de preguntes retrospectives; però aquestes presenten els seus propis problemes...

temps i detectar casos atípics o incoherències.

Des del punt de vista longitudinal, l'examen de la variabilitat entre i dintre subjectes de les variables d'interès és un pas lògic previ a la modelització. Quan és l'evolució intra-subjecte allò que centra el nostre interès, les possibilitats són diferents en funció del caràcter quantitatiu o qualitatiu de la nostra variable dependent.

En el cas d'una variable resposta mesurada al nivell d'interval, la tècnica exploratòria més habitual és l'anomenat *spaghetti plot*. Construir el gràfic de l'evolució d'una sub-mostra de subjectes en la variable dependent en funció de la mètrica de temps que haguem triat (per exemple, al llarg de les *x* rondes del panel), i també per a diferents nivells d'algunes covariables, resulta útil per generar hipòtesis entorn de la relació funcional que adopta l'evolució de la variable en el temps, el grau de variabilitat entorn de la mateixa i l'existència de trajectòries divergents.

Quan la nostra variable dependent és categòrica, les possibilitats d'anàlisi exploratòria estan menys estandarditzades. Una possibilitat equivalent a l'*spaghetti plot* és el *heat map,* també anomenat (amb petites diferències) *lasagna plot.* En el *heat map* la variació intraindividual en les categories de la variable dependent ve representada per un canvi de color en la trajectòria del subjecte (variable categòrica) o un canvi en la gradació del color (variable ordinal).[6]

En la nostra experiència, el *heat map* pot resultar poc aclaridor si no s'ordenen els individus d'acord amb alguna covariable significativament associada a la dependent. La perspectiva descriptiva oferida pel *heat map* es combina freqüentment amb la tècnica multivariada del *clustering,* que, a través de diverses variants, permet construir tipologies de trajectòries. Un exemple pràctic d'aquest procediment i de les altres possibilitats descriptives esmentades fins ara pot trobar-se en els exemples pràctics que acompanyen aquesta Guia.

Moltes altres possibilitats descriptives sorgeixen en funció dels aspectes específics del problema que ens estem plantejant. En els exemples pràctics hi ha exemples d'estratègies descriptives aplicades a diferents situacions.

2.2. Classificació de les tècniques estadístiques per a dades longitudinals

Som conscients que l'intent de classificar les tècniques estadístiques que s'ofereixen a l'investigador o investigadora a l'hora d'analitzar longitudinalment les dades longitudinals és, si més no, agosarat. La classificació que oferim ha estat elaborada d'acord amb els criteris següents:

1. Hem prioritzat la idea d'exhaustivitat per sobre de la d'especificitat. L'exposició detallada de moltes de les tècniques aquí tan sols esbossades, ocupa manuals voluminosos que no som capaços de substituir, ni ho pretenem. El propòsit és oferir a l'investigador o investigadora en ciències socials un plànol de les possibilitats fonamentals perquè pugui decidir en funció de les necessitats i els pressupòsits de la seva recerca. En el desenvolupament d'aquesta, requerirà una ajuda molt més detallada que la que aquesta Guia li pot oferir, encara que la bibliografia citada li pot servir de referent.

[6] Evidentment, el *heat map* també pot emprar-se amb una variable mesurada al nivell d'interval.

La voluntat d'exhaustivitat ha de ser immediatament matisada. Hem preferit excloure alguns desenvolupaments que per ara són minoritaris o amb els quals ens sentim particularment insegurs. No fem esment dels mètodes d'anàlisi de sèries temporals, ni de les seves extensions per als panels horitzontals. Tampoc no fem cap referència a aproximacions bayesianes. D'altra banda, només considerem les aproximacions no paramètriques[7] en relació amb l'anàlisi de patrons, i no pas en relació amb les aproximacions orientades a la variable.

L'anàlisi de correspondències (AC) i l'anàlisi de correspondències multiples (ACM), soles o amb combinació amb tècniques de cluster, també s'han fet servir per a l'anàlisi de dades longitudinals (Greenacre i Blasius, 1994, p. 231-280; Taris, 2000, p. 132-138). Tanmateix no sembla que aquesta opció estigui tenint continuïtat, amb la possible excepció de l'anàlisi de taules quadrades (Greenacre, 2008, p. 225-233), és a dir, en el context de l'anàlisi longitudinal, l'anàlisi de la matriu de probabilitats de transició dels individus entre l'estat *t* i l'estat *t*+1.

2. Tota classificació, d'altra banda, s'ha d'elaborar d'acord amb un o més criteris. En aquest cas n'hi ha molts de possibles i no sempre hem vist clar com es podien combinar.

Per això, el principi fonamental ha estat classificar d'acord amb allò distintiu que cada tècnica ofereix a l'investigador o investigadora. L'elecció de la tècnica estadística depèn del problema plantejat a la recerca.[8] Cada tècnica respon a un determinat tipus de preguntes o, si més no, les respon de forma diferent. Aquesta és la informació bàsica que cal.

Per aquesta raó hem distingit en primer lloc entre tècniques d'*anàlisi de patrons* i tècniques *orientades a la variable*. Bona part de la bibliografia ni tan sols no considera la primera possibilitat i se centra exclusivament en els enfocaments orientats a la variable. Per contra, nosaltres hem experimentat repetidament com en sociologia els investigadors i investigadores adopten intuïtivament la perspectiva de l'anàlisi de patrons.

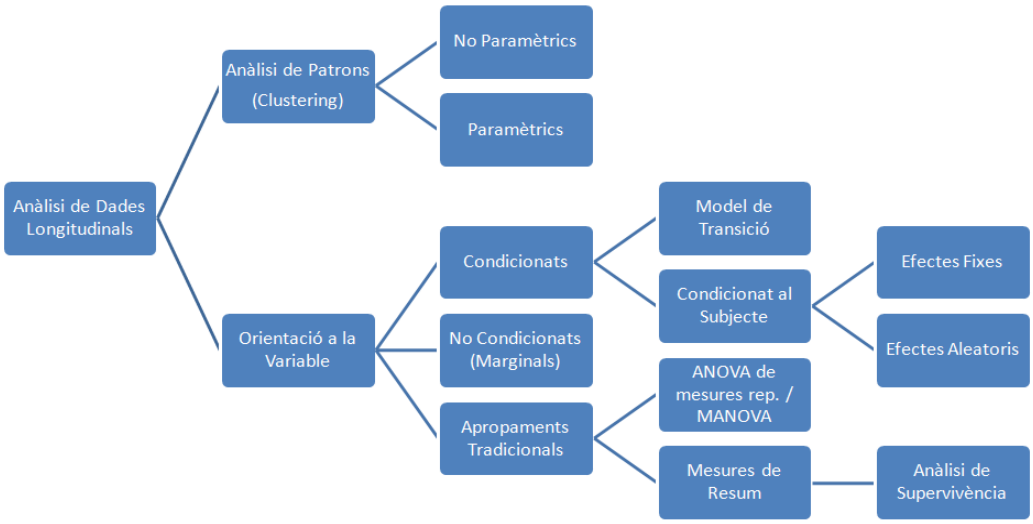
Hem combinat aquest criteris amb altres, que apareixen explícitament o implícita al llarg del text: 1) el nivell de mesura de la variable dependent (quantitativa, recompte, categòrica); 2) el caràcter paramètric o no paramètric de la tècnica; i 3) l'ús o no de variables latents (no observades).

Un parell d'aclariments per evitar confusions. Les tècniques estadístiques "per a dades longitudinals" rarament són adequades *exclusivament* per a dades longitudinals. Per això no ha de resultar estrany veure-les aplicades a altres tipus de dades. D'altra banda, com esdevindrà notori després, des d'un punt de vista matemàtic existeixen connexions entre moltes tècniques i sovint resulta possible combinar-les. En darrera instància, l'únic que existeix és la modelització de les dades, adaptada a les especificitats de cada

[7] Entenem per aproximacions paramètriques aquelles que assumeixen que les dades segueixen una determinada distribució de probabilitat els paràmetres de la qual pretenem estimar.
[8] La pràctica ens ensenya que aquest propòsit no s'aconsegueix completament en la realitat, ja que la recerca quantitativa exigeix preguntes comparativament simples i a la fi sempre es produeix un *trade-off* entre la pregunta original, i la manera en què finalment es planteja per tal de poder-la contrastar empíricament –ja sigui per la complexitat del model inicial o per limitacions de les dades disponibles.

problema. Això no impedeix tanmateix que es puguin fer distincions útils a l'hora d'orientar-se sobre l'estratègia d'anàlisi adequada en relació amb problemes concrets.

Figura 1. Classificació de les tècniques estadístiques d'anàlisi longitudinal



Font: elaboració pròpia

Orientació a l'anàlisi de patrons i orientació a la variable

Aquesta distinció fou explícitament introduïda per Bergman i Magnusson (1997) com a "perspectiva orientada a la persona" (person oriented approach) tot oposant-la a la "perspectiva orientada a la variable" (variable oriented approach).[9] També apareix amb un nivell d'elaboració menor en Taris (2000, p. 121), com a mode d'anàlisi centrat en la trajectòria vs. mode d'anàlisi centrat en l'esdeveniment.

De manera informal, entenem per tècniques orientades a l'anàlisi de patrons aquelles que estudien el vector de valors de l'individu en la variable dependent com una unitat primària d'informació, en lloc de descompondre'l en les variables que el constitueixen (Magnusson, Bergman i Rudinger, 1994, p. 19).

En canvi, en els mètodes orientats a la variable, aquesta és la veritable unitat conceptual i d'anàlisi. L'ènfasi es posa en la identificació de les relacions existents entre variables, i es suposa que les associacions trobades es mantenen per a tots els individus (Collins i Lanza, 2010, p. 8).

[9] Nosaltres hem preferit utilitzar l'expressió *anàlisi de patrons* en lloc de *perspectiva orientada a la persona* perquè la primera remet a un plantejament d'anàlisi estadística, mentre que la segona ho fa a un plantejament teòric.

Aquesta distinció no es considera rellevant en altres disciplines, per exemple en ciències de la salut, on la perspectiva dominant és clarament l'orientada a la variable. Però no ocorre així en les ciències socials. És per aquesta raó que considerem pertinent incloure l'anàlisi de patrons entre les possibilitats d'anàlisi de dades longitudinals, encara que és frequentment ignorada per la majoria de manuals.

L'adscripció a l'anàlisi de patrons o alternativament a l'orientat a la variable pot condicionar aspectes com ara: el format adequat de les dades, el tractament de la pèrdua d'informació i el tipus de predictors que es poden considerar (constants o variables amb el temps).

2.2.1. Tècniques estadístiques d'anàlisi de patrons

Encara que evidentment es pot recórrer a l'anàlisi de patrons sense subscriure els principis de la perspectiva orientada a la persona, hem considerat pertinent incloure més informació sobre aquesta perquè il□lumina alguns supòsits de l'anàlisi de patrons.

La perspectiva orientada a la persona està construïda sobre els principis de: 1) especificitat individual, 2) complexitat de les interaccions entre factors, 3) existència de diferències interindividuals en el canvi intraindividual, 4) estudi dels patrons com a mesura més adequada del canvi, 5) holisme i 6) existència d'un nombre limitat de classes de patrons –almenys a un nivell d'abstracció determinat (Sterba i Bauer, 2010, p. 240).

Aquest paradigma critica el plantejament de les tècniques orientades a la variable, basat en una descomposició analítica dels factors que intervenen en un procés de desenvolupament i el seu examen "dos a dos", procediment que destrueix la combinació única que resulta d'interaccions complexes entre factors diversos al nivell individual.

La perspectiva orientada a la persona condueix de forma natural, a l'hora de l'anàlisi estadística de les dades, a l'anàlisi de patrons (Bergman, Magnusson i El-Khouri, 2003).[10] Tot i l'ènfasi en l'especificitat individual, l'enfocament orientat a la persona no renuncia a identificar lleis de tipus general. La manera d'aconseguir-ho dintre d'aquest marc és identificar categories d'individus que presenten patrons similars en les seves característiques individuals (Collins i Lanza, 2010, p. 8).

En un context més específicament sociològic, la insatisfacció amb el "model de causalitat lineal" de relacions entre les variables, tal com apareix en els models de regressió estàndard i el seu caràcter "descontextualitzador" al nivell espacial, temporal i fins i tot social, fou un dels motius al□legats per Abbott per introduir l'anàlisi de seqüències i l'*òptimal matching* per abordar l'anàlisi longitudinal en les ciències socials (Blanchard, Buhlmann i Gauthier, 2012).

Les tècniques d'anàlisi de patrons són, doncs, essencialment, tècniques de classificació o clusterització dels individus. Quan es treballa sobre dades longitudinals, això es tradueix en una classificació de les trajectòries individuals possibles en relació amb el fenomen d'interès, i entre les quals els individus no poden transitar.

En un segon moment, les classes de trajectòries obtingudes es poden caracteritzar de

[10] Una revisió recent va mostrar que un 85% dels articles d'una importat revista en aquest àmbit utilitzaven tècniques classificatòries (Sterba i Bauer, 2010, p. 251).

forma descriptiva, probabilística, o bé es poden fer servir models de regressió o anàlisi discriminant per tal de clarificar els factors associats a una determinada classe de trajectòria.

De forma general, les tècniques d'anàlisi de clúster es poden dividir en paramètriques i no paramètriques (Mallapragada, Jin i Jain, 2010, p. 334).

2.2.1.1. Clusterització no paramètrica

Les tècniques de clúster no paramètriques inclouen algoritmes de clusterització populars en l'àmbit de les ciències socials com ara la classificació jeràrquica (*hierarchical cluster analysis*) o el *k-means*. Amb diverses variants, els algoritmes de classificació no paramètrica poden utilitzar-se per agrupar individus a partir tant de variables quantitatives com qualitatives, tant si aquestes són observades o factors latents –com ocorre quan es realitza clusterització sobre els factors d'una anàlisi de correspondències.

L'estratègia de Verd i López-Andreu (2012) en el seu estudi sobre trajectòries laborals amb dades del mateix PaD fou una mica diferent, ja que van fer la clusterització a partir d'indicadors resum de la trajectòria de l'individu (per exemple freqüència de temporalitat, inactivitat i atur, realització de formació no reglada, etc.). En qualsevol cas, l'objectiu és sempre la reducció del nombre inassumible de trajectòries individuals original a una tipologia de trajectòries parsimoniosa i teòricament interpretable.

Les tècniques de clúster no paramètriques agrupen els individus entre si en funció de la seva "proximitat" d'acord amb una definició específica de distància. L'*optimal matching* és una manera particularment interessant de definir les distàncies entre individus, emprada originalment en el camp de les ciències biològiques i importada amb èxit a les ciències socials per Abbott i Forrest (1986). Una aplicació de l'*optimal matching* es troba en el primer exemple d'aquesta Guia.

2.2.1.2. Clusterització paramètrica

Al llarg dels anys noranta del segle XX van començar a popularitzar-se tècniques de clúster "paramètric". A diferència de les anteriors, realitzen una sèrie de supòsits "forts" sobre la distribució de la població de la qual s'han extret les dades i els seus paràmetres. Tanmateix, comparteixen amb la seva contrapart no paramètrica el supòsit que la població analitzada està dividida en classes a les quals corresponen comportaments diferents en relació amb la variable o variables d'interès. Aquestes classes constitueixen una variable categòrica no observada (variable latent) el nombre de categories de la qual (i el valor dels paràmetres associats a cadascuna) tracten de determinar aquests models.

El fet que els clústers es determinin de forma paramètrica ofereix una sèrie d'avantatges, com ara l'existència de criteris més clars a l'hora de determinar el nombre de clústers adequat (*Akaike Information Criterion* –AIC–, *Bayesian information criterion* –BIC–); o bé el fet que l'adscripció d'un individu a un clúster no és dicotòmica (pertany/no pertany) sinó *probabilística*.

L'anàlisi de classes latents (*latent class analysis*) és una tècnica de clusterització no específicament pensada per a dades longitudinals. Tanmateix, pot ser aplicada a dades

de tipus longitudinal (en principi, categòriques), i llavors s'anomena anàlisi de classes latents de mesures repetides (*Repeated Measures Latent Class Analysis* –RMLTA– o anàlisi de classes latents longitudinal) (Collins, 2006, p. 22). El resultat del model, igual que en un clúster no paramètric, és l'agrupació de les trajectòries individuals en *k* categories i, si s'hi introdueixen covariables, s'obté una estimació del grau amb què prediuen la pertinença de l'individu a un determinat clúster.

Tècniques equivalents a l'RMLTA per a variable dependent quantitativa són els models de classes de creixement latents (*latent class growth modeling*) (Andruff, Carraro, Thompson, Gaudreau i Louvet, 2009) i la seva extensió, els models de creixement mixturats (*growth mixture modeling).* Aquests models suposen que el patró de creixement dels individus d'una població en la variable dependent és diferent per a diferents subgrups, i són capaços d'estimar el nombre d'aquests i els paràmetres que en caracteritzen la trajectòria de creixement. Aquestes tècniques comparteixen amb l'RMLTA el fet que cada classe latent és associada amb un vector de valors "característic" en la variable resposta (Collins i Lanza, 2010, p. 186-187) –i no amb cada component del vector per separat. En canvi, l'anàlisi de transicions latent (*latent transition analysis*), que està estretament associat a l'RMLTA, no modelitza el vector de mesures sencer de l'individu, sinó esdeveniment a esdeveniment, per tal d'estimar les probabilitats de transició de l'individu entre estats latents en dos moments consecutius del temps. Per aquesta raó, no compliria el criteri d'anàlisi de patrons tal com l'hem enunciat al principi.

2.2.2. Tècniques estadístiques orientades a la variable

Les tècniques estadístiques orientades a la variable modelitzen la variable dependent en cada moment del temps, en lloc de considerar el conjunt de la trajectòria per a un individu. S'orienten a identificar la naturalesa de les relacions existents entre les variables considerades, que s'assumeix que apliquen d'igual manera a tots els individus.

Des d'aquesta perspectiva, la trajectòria individual apareix no tant com una configuració específica de l'individu, sinó com l'efecte contingent d'una sèrie de factors (variables) que la determinen en cada moment. D'aquesta manera, la perspectiva orientada a la variable afavoreix, en major mesura que l'anàlisi de patrons, una lògica de la *intervenció* concreta que, tot introduint una modificació en un o més dels agents causals permeti, per exemple, "redreçar" una trajectòria considerada indesitjable.

La correlació entre els residus i l'ordenació en el temps de les mesures

Des del punt de vista de la modelització estadística, el gran problema que presenten les dades longitudinals és la dependència o correlació existent entre les residus de les mesures corresponents a una mateixa unitat d'anàlisi o individu.[11] Aquest fet viola un dels supòsits fonamentals de l'anàlisi de regressió i pot donar lloc a estimacions errònies dels coeficients i del seu error estàndard.

Un segon tret característic de les dades longitudinals és que les observacions d'un individu estan ordenades en el temps i no són necessàriament intercanviables. Això possi-

[11] Atès que la dependència de les dades no es presenta exclusivament en dades longitudinals (de fet ocorre sempre que les mesures estan agrupades en unitats més grans, que anomenem *conglomerats*), els models descrits aquí en general poden aplicar-se a estructures de dades diferents de la longitudinal.

bilita algunes solucions de modelització específiques per a dades longitudinals (Rabe-Hesketh i Skrondal, 2012a, p. 227-228).

La classificació que presentem s'organitza bàsicament en funció de les solucions que les diferents famílies de models donen a aquestes qüestions. La contrapartida és que cada família respon a un determinat tipus de preguntes de recerca. L'investigador o investigadora ha de ser conscient d'aquest fet a l'hora d'inclinar-se per una aproximació o una altra. D'altra banda, cal ser conscients que no estem davant de compartiments estancs, sinó que es poden combinar de diverses maneres.

Models condicionals i no condicionals

Deixant de banda les aproximacions tradicionals, sembla existir consens en la bibliografia quant a la classificació dels models utilitzats actualment per a l'anàlisi de dades longitudinals. Cal fer una primera distinció entre models condicionals i no condicionals (*unconditional models* o *marginal models; population averaged models*) (Bergsma, CroonMarcel i Hagenaars, 2009, p. 3; Molenberghs i Verbeke, 2005, p. 47-51; Rabe-Hesketh i Skrondal, 2012a, p. 227-229).

Els models condicionals estan explícitament interessats a esbrinar la natura de la dependència existent entre els residus d'un mateix subjecte. Per als models no condicionals o marginals, en canvi, el mecanisme que genera la dependència de les dades a nivell intraindividual no resulta en si mateix d'interès. Es tracta més aviat d'un inconvenient que cal tenir en compte a l'hora de realitzar les estimacions (Bergsma *et al.*, 2009, p. 3).

Dintre dels models condicionals, els models de transició (*transition models*, *dynamic models*) i els models condicionals al subjecte (*subject-specific models*) difereixen en la manera en què intenten explicar la dependència entre les dades.

Finalment, dintre els models condicionals al subjecte, els models d'efectes aleatoris i els models d'efectes fixos (*random effects models*, *fixed effects models*) difereixen en la manera en què controlen els efectes del subjecte.

Tant els models condicionals com els no condicionals s'han aconseguit ampliar (amb matisos) per poder modelitzar qualsevol tipus de variable dependent entre les unificades pel model lineal generalitzat (*Generalized Linear Model*), i per tant permeten modelitzar variables tant de tipus quantitatiu com categòric així com també recomptes.

2.2.2.1. Enfocaments tradicionals

Els primers mètodes per analitzar dades de tipus longitudinal van basar-se en adaptacions de l'anàlisi de variància a l'existència de mesures repetides.

En l'anàlisi univariat de variància de mesures repetides (*repeated measures ANOVA*) s'introdueix l'efecte de l'individu com un factor aleatori en el model. Aquest factor representa l'efecte agregat de tots els factors no mesurats específics de cada individu i, d'aquesta manera, es controla la correlació existent entre les seves mesures.

L'anàlisi multivariat de la variància (MANOVA), en canvi, és una veritable tècnica multi-

variant en què les diferents mesures d'un mateix individu al llarg del temps són tractades com un únic vector. Des del punt de vista dels supòsits és menys restrictiu que l'ANOVA de mesures repetides.

Una tercera aproximació al problema fou reduir el vector multivariant de mesures d'un individu a una única mesura de resum (*summary values*) –per exemple, l'àrea sota la corba de la trajectòria d'un individu. Un cop obtingut això, es podien aplicar els models ANOVA tradicionals.

Encara que aquestes tècniques possibilitaven l'anàlisi de dades longitudinal, resultaven insatisfactòries en diversos sentits. Els models de l'ANOVA de mesures repetides i el MANOVA incloïen supòsits massa restrictius, i presentaven dificultats en presència de dades no equilibrades o mesurades a intervals irregulars –a més de no resultar útils quan la variable dependent era de tipus categòric.

Quant a l'enfocament de la "mesura de resum", implicava inevitablement una pèrdua d'informació. Individus amb un vector multidimensional diferent podien originar un mateix valor resum, mentre que tampoc no permetia estudiar l'efecte de les variables dependents del temps.

Tot això va estimular la recerca de tècniques d'anàlisi longitudinal més flexibles (Fitzmaurice i Molenberghs, 2009, p. 4-7).

L' anàlisi de supervivència

A diferència de les tècniques anteriors, les anàlisis de supervivència continuen plenament vigents. Les anàlisis de supervivència es caracteritzen perquè modelitzen una variable dependent molt específica: el temps transcorregut fins que el subjecte experimenta un canvi qualitatiu d'estat.

Una manera d'explicar la pervivència de l'anàlisi de supervivència és relacionar-lo amb l'aproximació de les "mesures de resum" suara explicada. Hem vist que aquesta resulta insatisfactòria ja que implica sempre algun tipus de pèrdua d'informació.

Doncs bé, en el cas de les anàlisis de supervivència, el temps fins a l'ocurrència de l'esdeveniment constitueix un resum *suficient* de la història de l'individu. En realitat, no existeixen mesures repetides per subjecte, sinó una única mesura.[12] L'anàlisi de supervivència ha hagut de desenvolupar tècniques específiques per modelitzar la funció del temps i l'existència de casos censurats (és a dir, casos dels quals ignorem el desenllaç respecte a l'esdeveniment d'interès), però no es troba d'entrada el problema de mesures correlacionades intrasubjecte.[13]

Inicialment les anàlisis de supervivència es van aplicar al temps fins a l'ocurrència d'un event dicotòmic. L'evolució de les tècniques estadístiques ha tornat possible la modelit-

[12] Això es veu clarament en l'anàlisi de riscos a temps discret que constitueix el darrer exemple d'aquesta guia: per elaborar la taula de vida només ens cal la darrera observació de la llar, mentre que la caiguda o no de la llar es modelitza mitjançant una regressió logística estàndard.
[13] Llevat, naturalment, que l'esdeveniment d'interès hagi estat definit com a repetible. En aquest cas, la modelització també haurà de tenir en compte aquest problema.

zació de situacions més complexes.

2.2.2.2. Models condicionals al subjecte

Els models condicionals al subjecte realitzen l'estimació dels coeficients del model condicionals a aquell o, si es vol, introdueixen els efectes del subjecte en el model per tal de controlar-los. Els models d'efectes aleatoris introdueixen l'efecte del subjecte en la part aleatòria del model, és a dir, en el terme d'error. Els models d'efectes fixos, en canvi, introdueixen els efectes del subjecte en la part fixa del model, això és, com una variable explicativa més.

Models d'efectes aleatoris

Els models d'efectes aleatoris, o models multinivell, o models jeràrquics (*multilevel models*, *hierarchical models*) es fan servir quan les dades de la mostra no corresponen a un esquema de mostreig aleatori simple, sinó que estan agrupades en conglomerats, fet que provoca (normalment) correlació entre les unitats d'un mateix conglomerat (per exemple seccions censals seleccionades en la primera etapa i individus seleccionats dins d'aquestes en la segona etapa).

Des d'aquesta perspectiva, les dades longitudinals poden contemplar-se com dades agrupades (conglomerats), en què el primer nivell són les diverses mesures i el segon l'individu al qual pertanyen. La introducció d'efectes aleatoris permetrà controlar l'efecte que té sobre les mesures el fet de pertànyer al mateix conglomerat (individu).

En els models d'efectes aleatoris, els efectes de l'individu s'introdueixen en la part aleatòria o residual del model. Més concretament, la part aleatòria del model es subdivideix en dues: la corresponent al factor o factors aleatoris específics del subjecte, i la part residual pròpiament dita.

La part fixa (és a dir, les variables explicatives) del model d'efectes aleatoris permet explicar la tendència mitjana de les dades, mentre que els factors aleatoris expliquen la variabilitat individual en relació a aquesta tendència mitjana, i permeten així un millor ajust del model a les dades.

Les aplicacions habituals d'aquest tipus de models són el desenvolupament psicològic, el creixement físic o l'aprenentatge d'una persona, en què tant la naturalesa com les raons de la variabilitat són el major tema d'interès. Els models de corbes de creixement són un cas especial dels models d'efectes aleatoris. En aquest enfocament multinivell a l'anàlisi de dades longitudinal, el focus està en la modelització del creixement o decreixement per representar les trajectòries de creixement individual.

Models d'efectes fixos (fixed effects models)

A diferència dels models d'efectes aleatoris, els models d'efectes fixos controlen l'efecte de l'heterogeneitat interindividual introduint-la en la part fixa del model, com una variable explicativa més.

Mitjançant aquesta estratègia, els models d'efectes fixos es desempalleguen completa-

ment de la variació entre subjectes i es concentren exclusivament en l'explicació de la variança intraindividual. Això fa que l'error de l'estimació augmenti considerablement en relació amb els models d'efectes aleatoris, que combinen les dues fonts de variació interindividual i intraindividual. Aquesta pèrdua de precisió es veu compensada per l'eliminació de l'heterogeneïtat no observada que molt probablement conté la variança entre subjectes, la qual cosa permet obtenir estimacions no esbiaxiades (Allison, 2005, p. 4).

Una limitació important dels models d'efectes fixos és que no poden estimar els coeficients d'una variable quan la variabilitat intraindividual d'aquesta tendeix a zero (variables constants en el temps), com ocorre amb el sexe, el lloc de naixement o fins i tot el nivell d'estudis. Encara que es poden introduir interaccions entre variables constants, i d'altres que varien amb el temps, els models d'efectes fixos en el context d'estudis observacionals es fan servir sobretot per investigar els efectes de covariables dependents del temps (Allison, 2005, p. 4).

Això, juntament amb la pèrdua potencial d'eficiència, ha condicionat una infrautilització dels models d'efectes fixos en l'àmbit de la sociologia (Halaby, 2004), mentre que són en canvi l'opció de preferència en l'àmbit de l'economia.

2.2.2.3. Models de transició o dinàmics (*transition models, dynamic models*)

En els models de transició, continuem interessats en la natura de la dependència existent entre observacions. Però, en aquest cas, no intentem explicar aquesta dependència a partir de particularitats individuals, sinó a partir de valors anteriors de la mateixa variable dependent.

En aquests models, la variable dependent en el moment *t*-1 (i opcionalment *t*-2, *t*-3...) és utilitzada juntament amb la resta de covariables del model per explicar el valor de la variable dependent en el moment *t*.

En utilitzar models dinàmics, estem suposant que la conducta, les creences o les actituds prèvies estan determinant la conducta, les creences o les actituds posteriors. Per exemple, els salaris de fa un any poden afectar els salaris actuals. Dit d'un altra manera, el procés estudiat mostra "dependència de l'estat" (*state dependence*). En canvi, quan utilitzem models condicionals al subjecte, estem suposant que el procés mostra "dependència del subjecte", és a dir són les diferències individuals (en variables no observades) les que expliquen les diferències interindividuals en el procés.

Quan es combinen amb models condicionals al subjecte, ja siguin d'efectes fixos o aleatoris, els models dinàmics permeten decidir entre les dues explicacions alternatives de la dependència: causada per dependència de l'estat anterior, o per diferències interindividuals (Skrondal i Rabe-Hesketh, 2013).

2.2.2.4. Models no condicionals, marginals o *population averaged models*

A diferència dels models condicionals, els models no condicionals no s'interessen per la natura de la dependència entre les observacions d'un mateix subjecte. Aquesta és contemplada més aviat com una molèstia que cal tenir en compte a l'hora de realitzar les estimacions, però l'interès dels models marginals es concentra en la tendència general de les dades (Bergsma *et al.*, 2009, p. 3).

Els models marginals sovint es fan servir en biomedicina per analitzar dades procedents d'assajos clínics experimentals, a fi d'estimar els efectes mitjans d'un tractament. En aquest context, se suposa que l'efecte de les diferències individuals ja ha estat controlat per l'assignació aleatòria dels subjectes als diferents nivells de tractament, i les diferències individuals tenen un interès secundari (Rabe-Hesketh i Skrondal, 2012a, p. 229).

Des del punt de vista de la modelització, mentre que en els models d'efectes aleatoris el terme d'error es divideix entre un o més factors aleatoris i l'error residual, en els models marginals la dependència intraindividual es tracta especificant una estructura determinada de correlacions per a l'error total.

Un aspecte que s'ha de tenir en compte en cas d'utilitzar models marginals és que les ocasions de mesura han de ser aproximadament "simultànies" per a tots els individus.

Estimacions condicionals i no condicionals

Els models condicionals al subjecte, ja siguin fixos o aleatoris, estimen l'efecte de les variables independents condicionals a les característiques del subjecte. En canvi, els models marginals estimen l'efecte "mitjà" de les variables independents sobre la variable resposta al nivell poblacional.

Les estimacions dels coeficients de regressió poden resultar molt diferents segons si s'adopta un model condicional al subjecte o un de no condicional. Això resulta especialment cert en el cas de dades categòriques i models no lineals (Bergsma *et al.*, 2009, p. 3).

En realitat, cada tipus d'estimació respon a una pregunta diferent. Si estem interessats a predir el valor o l'evolució en la variable dependent d'un individu en concret, és millor utilitzar els coeficients dels models condicionats al subjecte. En canvi, si allò que ens interessa és predir l'evolució poblacional o l'efecte poblacional d'una determinada covariable, són millors les estimacions poblacionals.

Així doncs, pot semblar que en ciències socials sempre resultaran preferibles els models marginals. Si per exemple volem estimar l'efecte d'una campanya contra l'ús d'una droga, normalment estarem més interessats a veure com ha evolucionat el consum de la substància per al conjunt de la població (i les variables que hi han influït), que no pas com evoluciona el consum al nivell individual.

Ara bé, cal ser conscients que amb les estimacions del model marginal, romanem ignorants respecte del mecanisme que provoca el canvi al nivell individual, i que és el veritable origen del canvi poblacional: l'evolució a nivell individual no pot inferir-se de l'evolució a nivell poblacional, ja que l'evolució mitjana per al conjunt de la població pot ser el resultat agregat de trajectòries molt diferents. Els models marginals es limiten a estimar efectes al nivell poblacional, sense informar sobre els processos individuals subjacents (Bergsma *et al.*, 2009, p. 108).

Per la seva banda, en els models de transició les estimacions dels coeficients de regressió es realitzen condicionades al valor o valors anteriors de la variable, entre altres factors. Aquest fet és criticat per alguns autors, ja que en aquestes condicions esdevé difí-

cil interpretar què representen aquests coeficients (Molenberghs i Verbeke, 2005, p. 49-50).

2.3. Programari estadístic per analitzar dades de panel

Existeixen programes excel·lents especialitzats en l'ajust de determinats tipus de models, per exemple MLwin per a models d'efectes aleatoris, Mplus per a models de variable latent o LatentGold per al cas d'anàlisi de classes latents. Gairebé sempre aquesta mena de programari permet ajustar models d'altres tipus, per la qual cosa la seva utilitat va més enllà. El programa Sleipner consta d'una col·lecció de mòduls amb eines estadístiques per a l'anàlisi de dades des del paradigma orientat a la persona.

De tota manera, creiem que per a introduir-se en les tècniques de recerca longitudinal són suficients els mòduls corresponents de paquets estadístics generalistes com ara SPSS, Stata, R o SAS. Nosaltres tenim les nostres preferències i, sense declarar-les, direm que tenen relació amb la facilitat d'interpretació tant de la sintaxi que permet ajustar els models, com de l'*output*.

Com és habitual, alguns dels programes esmentats són de programari lliure, mentre que d'altres són comercials i en moltes ocasions de preu inassumible, si no es disposa del recolzament d'una institució.

En relació al programari, el nostre missatge principal és que l'investigador o investigadora interessat en anàlisi longitudinal hauria de desenvolupar una actitud oberta i no inhibir-se d'explorar programes que li són desconeguts. En alguns la corba d'aprenentatge és lenta (per exemple R) mentre que d'altres (per exemple Stata) són inicialment més amigables i disposen d'una documentació excel·lent accessible directament des de la web.

3. Exemples pràctics amb dades del Panel de Desigualtats de Catalunya (PaD)

A continuació es desenvolupen tres exemples de tècniques d'anàlisi estadística per a dades longitudinals, aplicades a dades procedents d'un disseny de panel. En concret, corresponen a diverses submostres del Panel de Desigualtats Socials a Catalunya (PaD). Podeu accedir als fitxers de dades necessaris per seguir-los en aquest *enllaç*.

El PaD és un panel de llars (*household panel*) dut a terme amb periodicitat anual per la Fundació Jaume Bofill al llarg del període 2002-2012. La metodologia del PaD és similar a la del *British Household Panel Survey*, un dels estudis de panel de llars més reconeguts a escala internacional (Fundació Jaume Bofill, 2009, p. 9).

La mostra original (1.991 llars i 5.785 individus) era representativa del conjunt de la població catalana. Les llars es van seleccionar mitjançant un disseny mostral estratificat i multietàpic, amb mostreig sistemàtic de les unitats primàries (seccions censals) i mostreig aleatori simple de les unitats secundàries (adreces). Les seccions censals es van estratificar d'acord amb la província, la mida del municipi, el percentatge de classe treballadora i el percentatge de persones econòmicament actives (Fundació Jaume Bofill,

2009, p. 27-31).

Les províncies de Girona, Lleida i Tarragona es van sobrerepresentar per tal de garantir la significació estadística de les estimacions al nivell provincial. L'assignació final fou de 400 llars per província excepte a la de Barcelona, a la qual correspongueren les 791 llars restants (Fundació Jaume Bofill, 2009, p. 32).

A partir del 2006 i a causa del desgast mostral que posava en perill la representativitat de la mostra, es va procedir al remostreig de noves llars. D'aquesta manera, el 2009 3.266 llars i 9.544 individus havien format part del PaD en algun moment.

El PaD recull informació tant al nivell de llar com d'individus. Les dades sobre la llar es recullen mitjançant un qüestionari de llars, que és respost per la persona designada com a informant principal de la mateixa. Així mateix es recull informació sobre els individus a partir de 16 anys a través d'un qüestionari d'individus, o bé a través un qüestionari proxy, que és respost per l'informant principal de la llar en nom de la persona interessada, si aquesta es troba absent en el moment de realitzar l'entrevista. Tanmateix, el qüestionari proxy és una versió simplificada del qüestionari d'individus, fet que implica la pèrdua d'informació en moltes de les variables de tipus individual.

El qüestionari de llars recull informació sobre els membres de la llar, relacions familiars, treball domèstic, condicions de llar, pressupost familiar i consum. El qüestionari individual, per la seva banda, recull informació sobre els aspectes més determinants de la vida i el benestar de l'individu com ara l'educació, la situació respecte al mercat de treball, el treball reproductiu, els usos del temps, el nivell socioeconòmic, la salut, així com valors i actituds diversos.

Un tret diferencial del PaD és que el treball de camp ha estat dut a terme per personal directament contractat per la Fundació Jaume Bofill, contràriament a la majoria de panels a llars, en què aquesta tasca s'externalitza a companyies especialitzades.

El PaD ha sofert diversos canvis metodològics menors al llarg de la seva existència. L'aspecte més important és sens dubte el canvi en el mètode de recollida de dades. Des del 2002 i fins al 2009, les dades es van recollir mitjançant entrevista personal assistida per ordinador (CAPI, *Computer Assisted Personal Interviewing*); mentre que des del 2010 fins a la darrera ronda del PaD, el 2012, les entrevistes es van realitzar telefònicament, mitjançant el sistema CATI, *Computer Assisted Telephone Interviewing*.

Juntament amb aquesta Guia es faciliten els arxius de dades i de comandaments en el format del programa estadístic corresponent a cada cas, de manera que el lector interessat pugui reproduir els resultats si així ho desitja.

Les submostres del PaD utilitzades en els exemples contenen dades recollides mitjançant CAPI entre el 2002 i el 2009 –bàsicament per no disponibilitat de les tres darreres rondes en el moment de planificar aquesta Guia. Criteris de selecció més específics han estat aplicats (i explicitats) en cada cas. Cal aclarir que aquestes submostres no s'obtenen directament de la base de dades del PaD, sinó que han estat objecte d'algunes transformacions prèvies –especialment les que serveixen de base a l'anàlisi de supervivència.

De forma general, en aquests casos pràctics no hem adreçat el problema de la representativitat de la mostra, ni el caràcter clusteritzat de les dades al nivell de llar, aspectes que sens dubte caldria tenir en compte en un procés de recerca real.

3.1. Anàlisi de seqüències: trajectòries laborals

En aquest estudi de cas es mostra un exemple d'anàlisi i visualització de trajectòries laborals de les persones entrevistades al PaD en les onades del 2002-2006, la creació d'agrupacions de les persones en funció de les seves seqüències laborals i finalment la determinació de potencials variables explicatives d'aquestes agrupacions.

El programari utilitzat ha estat el paquet TraMiner de l'entorn R, Versió 1.8-6.

3.1.1. Plantejament de l'estudi

Qué són les trajectòries laborals?

Entenem per trajectòria laboral la seqüència d'estatus o situacions laborals d'una persona registrades per unitat de temps (Massoni, Olteanu i Rousset, 2009), és a dir, el conjunt de totes les situacions possibles respecte al mercat de treball en què es troba una persona al llarg d'un període d'estudi, en què cada situació es mesura en una unitat de temps, que pot ser el mes, el trimestre o l'any. Així, una persona pot haver estat durant tot el període d'estudi en una mateixa situació, per exemple "inactivitat", o "treballant amb contracte fix", mentre que també pot haver anat canviant la seva situació laboral, cada any, variant entre situacions d'"inactivitat" o "atur".

Per què és important analitzar les trajectòries laborals des d'una perspectiva longitudinal? L'estudi de les trajectòries laborals ens acosta a quina ha estat l'evolució d'una persona en el mercat de treball, és a dir la seva "biografia laboral". Concretament, podem analitzar quina ha estat l'estabilitat o inestabilitat de la persona quant a la seva situació laboral, aspectes que no es poden recollir quan s'analitza la situació laboral de la persona de manera transversal, en un moment específic del temps. En l'article de Zubiri-Rey (2008) es presenten amb més detall alguns exemples específics de la potencialitat d'analitzar les trajectòries laborals.

Objectiu de l'estudi

En aquest estudi de cas es mostra com es poden descriure les trajectòries laborals de les persones entrevistades al PaD en el període 2001-2006 des d'una perspectiva longitudinal. A més, es presenta com es poden agrupar aquestes trajectòries laborals i caracteritzar aquestes agrupacions d'acord amb altres variables potencialment explicatives, com per exemple, de tipus sociodemogràfic i laboral. Com a possibles determinants de les agrupacions s'han escollit les variables: sexe, edat a l'inici de l'estudi i nivell d'ingressos.

Els resultats d'aquest estudi de cas es poden contrastar amb els que presenten Verd i López-Andreu (2012), també amb dades del PaD per al mateix període, però on es consideren altres tècniques estadístiques alternatives.

Operacionalització de les trajectòries laborals

Per a estudiar les trajectòries laborals s'ha considerat la situació laboral de la persona en cadascuna de les rondes d'estudi, considerant cinc possibles situacions o estatus laborals: "inactiva", "aturada", "treballadora amb contracte temporal", "treballadora amb contracte fix" i "treballadora funcionària".

Segons Massoni, Olteanu i Rousset (2009), com més estatus laborals es defineixin, més acurat serà l'estudi de seqüències laborals. No obstant això, cal tenir present que la definició dels estatus o situacions laborals s'ha de basar en un model teòric previ.

Selecció de casos o base mostral

L'anàlisi de les seqüències laborals s'ha limitat a les persones entrevistades al PaD en les cinc primeres rondes (2002-2006) amb tota la informació disponible per a les variables d'estudi, és a dir, sense dades mancants, un aspecte que limita l'anàlisi de seqüències amb les tècniques estadístiques que mostrem a continuació. La mostra seleccionada inclou 1.055 persones amb 5.275 registres acumulats al llarg de les cinc rondes analitzades.

Variables d'estudi

A la taula 1 es resumeixen les variables d'estudi. Per a construir la variable d'interès, és a dir, les trajectòries laborals, es consideraran cinc variables que recullen les situacions o estatus laborals de la persona en cadascuna de les cinc rondes d'estudi. Aquestes variables són de tipus categòric nominal, amb cinc categories o estats possibles (Inactiu, Aturat, Temporal, Fix i Funcionari). Com a variables potencialment explicatives de les agrupacions de trajectòries laborals hem escollit dues variables de tipus continu (l'edat a la primera ronda i els ingressos a la darrera ronda), una variable categòrica nominal (el sexe de la persona entrevistada) i una variable categòrica ordinal (el nivell d'estudis a la primera ronda).

Cal notar que en aquest exemple hem simplificat les variables explicatives recollint l'observació d'una única ronda com a referència (la primera o la darrera). No obstant això, en el cas del nivell d'estudis, l'edat i els ingressos es podrien haver pres les variables explicatives mesurades en les cinc rondes, considerant que aquestes poden variar en el temps, és a dir, són variables "dependents del temps", com ho és la situació laboral.

Taula 1. Variables d'estudi

Tipus	Nom	Descripció	Valors
Dependent o resposta	**Estatus o situació laboral_1**	Indica la situació laboral de la persona a la primera ronda	1: Inactiu 2: Aturat 3: Temporal 4: Fix 5: Funcionari

	Estatus o situació laboral_2	Indica la situació laboral de la persona a la segona ronda	1: Inactiu 2: Aturat 3: Temporal 4: Fix 5: Funcionari
	Estatus o situació laboral_3	Indica la situació laboral de la persona a la tercera ronda	1: Inactiu 2: Aturat 3: Temporal 4: Fix 5: Funcionari
	Estatus o situació laboral_4	Indica la situació laboral de la persona a la quarta ronda	1: Inactiu 2: Aturat 3: Temporal 4: Fix 5: Funcionari
	Estatus o situació laboral_5	Indica la situació laboral de la persona a la cinquena ronda	1: Inactiu 2: Aturat 3: Temporal 4: Fix 5: Funcionari
Independents o explicatives	**Sexe**		1: Home 2: Dona
	Edat	Edat a l'inici del seguiment	Contínua
	Nivell d'estudis	Nivell d'estudis a la primera ronda	1: Primària o menys 2: 1r cicle secundària 3: 2n cicle secundària, no tècnic 4: 2n cicle secundària, tècnic 5: Postsecundària 6: Estudis superiors
	Ingressos	Ingressos a la cinquena ronda	Contínua

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

3.1.2. Tècniques estadístiques per a estudiar les trajectòries laborals

En aquest estudi de cas, es presenten tres aproximacions a l'anàlisi estadística de les trajectòries laborals mitjançant el programari estadístic d'accés lliure R.

Per a dur a terme aquestes anàlisis s'utilitzaran principalment les aplicacions del paquet TraMineR, desenvolupat per Gabadinho (2011a; 2011b), especialment a l'estudi de les trajectòries laborals amb un enfocament longitudinal.

Concretament, es farà un breu recorregut per les tècniques d'anàlisi descriptiva i l'anàlisi multivariant d'agrupació de les trajectòries. A més, es presenta una aplicació

per a l'estudi de les característiques de les diferents agrupacions de persones d'acord amb algunes variables socials d'interès com ara el sexe, l'edat, el nivell d'ingressos o el nivell d'estudis, mitjançant models de regressió logística binomial.

Format de la base de dades

El paquet TraMineR permet introduir les dades en diferents formats per identificar els estats o esdeveniments que poden tenir lloc per a un mateix subjecte d'estudi, en les unitats de temps corresponents. El format "Seqüència d'estats" (STS) és el format utilitzat per defecte amb TraMineR i és també el més intuïtiu i comú a l'hora de representar les seqüències, les dades hi són organitzades de manera que els estats successius d'un subjecte d'estudi es presenten en diferents columnes, corresponents a cadascuna de les unitats de temps. Aquest és el cas de les dades del nostre exemple, en què tenim cinc columnes que recullen els cinc estatus laborals de les persones entrevistades en les cinc rondes. Així doncs, en aquest format tindrem una única fila o registre per a cada individu. A banda de les variables d'estat a les cinc rondes, a continuació podem tenir altres variables d'interès, que poden haver estat mesurades en les cinc rondes, és a dir, que poden ser dependents del temps, o bé variables no dependents del temps. Com s'ha comentat en el subapartat Variables d'estudi, en aquest exemple ho simplificarem al cas de variables no dependents del temps.

El mateix paquet TraMineR porta implementada una funció que permet convertir les dades d'un format a un altre:

```
seqformat()                                                    (i)
```

Per a més informació sobre els diferents tipus de formats de dades disponibles, recomanem llegir el capítol 4 de Gabadinho (2011b).

3.1.3. Execució, interpretació de resultats i comentaris amb R

En primer lloc, crearem l'objecte que recull les dades d'anàlisi "dades". R permet llegir directament des de qualsevol paquet estadístic o base de dades (en aquest cas SPSS). No obstant això, prèviament s'ha d'especificar la llibreria `foreign`:

```
library(foreign)                                               (ii)
dades<-read.spss('Estudi_cas_1.sav')
```

A continuació podem definir aquest objecte "dades" com el nostre `data.frame` o fitxer de variables i observacions:

```
dades<-as.data.frame(dades)                                    (iii)
```

I finalment, per assegurar-nos que la lectura s'ha fet correctament podem utilitzar la instrucció `head ()`:

```
head(dades)                                                    (iv)
```

Quadre 1. Base de dades amb R: variables d'estudi i valors de les variables per a una mostra de persones entrevistades (n=6). Panel de Desigualtats (PaD). Catalunya, 2002-

2006

```
       id_ind status.1 status.2 status.3 status.4 status.5 sexe
1 111100010001  Inactiu   Aturat     Fix     Fix   Aturat Home
2 111100070002  Inactiu  Inactiu Temporal Temporal Temporal Dona
3 111100200001  Inactiu  Inactiu  Inactiu  Inactiu  Inactiu Dona
4 111101020002  Inactiu  Inactiu Temporal Temporal  Inactiu Dona
5 111101150002  Inactiu  Inactiu  Inactiu  Inactiu   Aturat Dona
6 111101620001     Fix     Fix     Fix Temporal  Inactiu Home

                              estudis edatinici ingressosfi
1                  'Primària o menys'        56    655.9118
2 'Segon cicle de secundària, tècnic'        49    537.0000
3                  'Primària o menys'        33      0.0000
4 'Segon cicle de secundària, tècnic'        46      0.0000
5 'Segon cicle de secundària, tècnic'        48    225.8524
6                 'Estudis superiors'        37   1200.0000
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Definició de les trajectòries laborals

A continuació, caldrà definir les seqüències o trajectòries laborals per a cada individu. Per això, s'ha d'indicar on es troben les variables de situació laboral a la base de dades, en el nostre cas, a les columnes de la 2 a la 6. (La primera es correspon amb l'identificador.) Abans, però, cal carregar la llibreria corresponent al TraMineR. Si no la tenim prèviament instal·lada en el nostre entorn R, ho haurem de fer a través de l'opció `packages-> install package(s)` del menú de R. Triarem el Cranmirror des del qual volem descarregar el TraMineR (per exemple, Denmark) i seleccionarem l'opció TraMineR.

```
library(TraMineR)                                                    (v)
dades.seq<-seqdef(dades,var=2:6)
```

Una vegada hem definit la seqüència o trajectòria laboral per a cada individu, el programa ens ho confirma detallant l'alfabet, és a dir els diferents estats possibles; en aquest cas cinc, i les seves etiquetes (Aturat, Fix, Funcionari, Inactiu i Temporal). També ens informa del nombre de seqüències laborals (1.055), corresponents a totes les persones entrevistades, i de la longitud de les seqüències laborals (5).

Quadre 2. Confirmació de la creació de les seqüències laborals i descripció dels estats possibles. Panel de Desigualtats (PaD). Catalunya, 2002-2006

```
 [>] 5 distinct states appear in the data:
     1 = Aturat
     2 = Fix
     3 = Funcionari
     4 = Inactiu
     5 = Temporal
 [>] state coding:
       [alphabet]  [label]    [long label]
```

```
     1  Aturat      Aturat     Aturat
     2  Fix        Fix       Fix
     3  Funcionari  Funcionari Funcionari
     4  Inactiu     Inactiu    Inactiu
     5  Temporal    Temporal   Temporal
 [>] 1055 sequences in the data set
 [>] min/max sequence length: 5/5
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

En el nostre cas tenim cinc tipus diferents d'estats, que es codifiquen i etiqueten amb els valors i noms de la base de dades original d'SPSS. Tenim en total 1.055 seqüències laborals, corresponents a 1.055 persones entrevistades a les cinc primeres rondes del PaD (2002-2006).

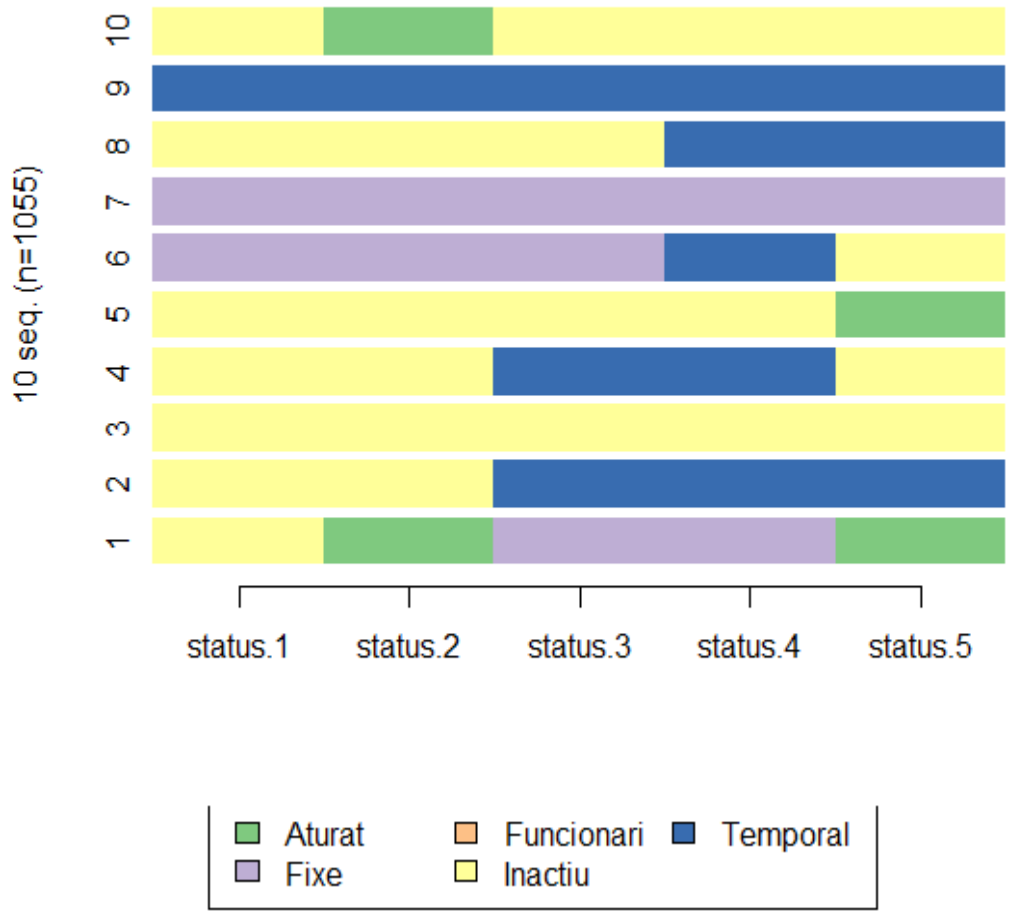
Una vegada s'han creat les trajectòries laborals, el TraMineR ofereix la possibilitat de fer diverses aproximacions a l'anàlisi estadística de les seqüències. A continuació, es presenten alguns exemples de les possibilitats d'anàlisi exploratòria o descriptiva de les dades.

Anàlisi descriptiva de les seqüències laborals

Començarem analitzant gràficament les deu primeres trajectòries (figura 2). La instrucció `(vi)` ens permet veure el gràfic de l'índex de trajectòries que ens mostra la successió individual d'estats laborals:

```
seqiplot(dades.seq,border=NA)                                    (vi)
```

Figura 2. Exemple de les trajectòries laborals de deu persones del Panel de Desigualtats (PaD)



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Podem veure com algunes persones han mantingut una trajectòria estable, mentre que d'altres han estat més variables. Per exemple, mentre que els individus 3, 7 i 9 s'han mantingut en situació inactiu, fix i temporal, respectivament, al llarg de les cinc rondes d'estudi, la resta ha anat variant entre períodes d'inactivitat, atur, treball fix i temporal.

En primer lloc, la instrucció `seqtab (vii)` ens permet obtenir les freqüències absolutes i els percentatges corresponents de les deu trajectòries més freqüents presents a la mostra:

```
seqtab(dades.seq, tlim=1:10)                                    (vii)
```

L'opció `tlim` ens permet fixar quantes trajectòries volem analitzar. En el nostre cas, les deu trajectòries laborals més freqüents són el 67,8% de totes combinacions d'estatus laboral representades a la mostra d'estudi (quadre 3).

La trajectòria més freqüent correspon a aquelles persones que han mantingut la situació de treball fix al llarg de les cinc rondes (39,72%), seguides de les que són funcionà-

ries (11,18%). A continuació, amb un 5,40%, troben les establement temporals i, ja amb menys d'un 4%, tota la resta de combinacions.

Quadre 3. Trajectòries laborals de les persones entrevistades al Panel de Desigualtats (PaD)

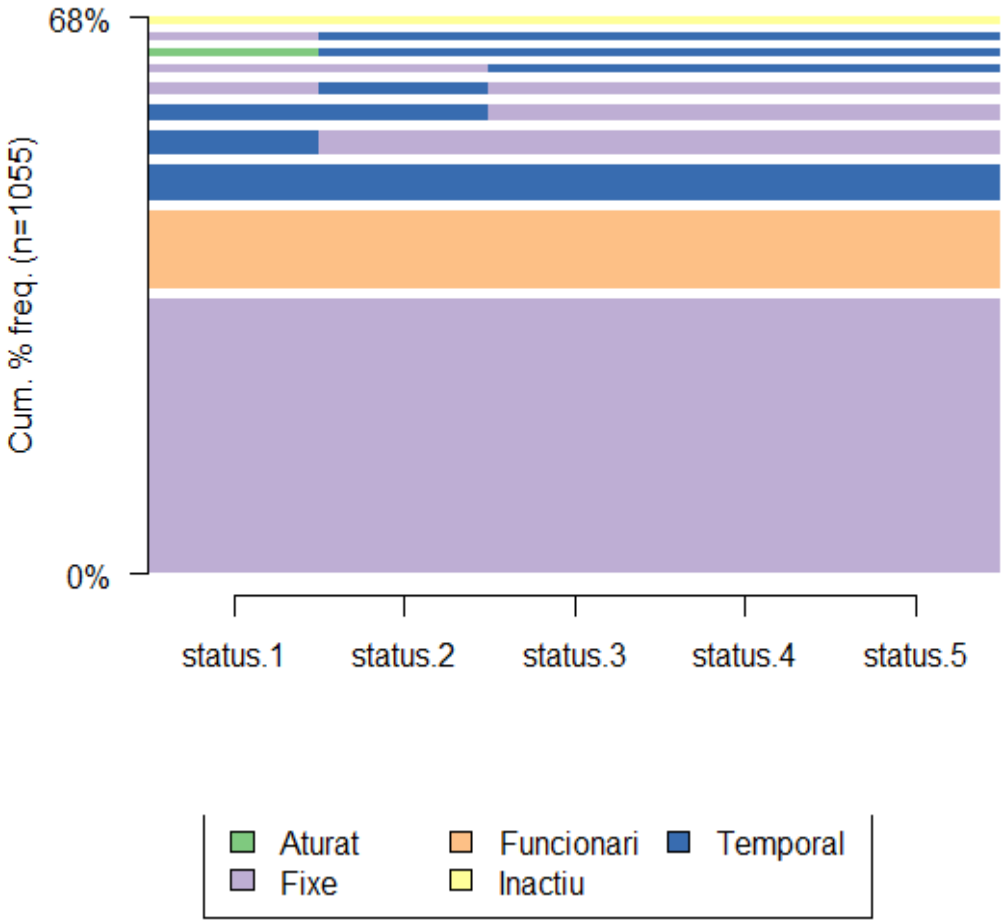
	Freq	Percent
Fix/5	419	39.72
Funcionari/5	118	11.18
Temporal/5	57	5.40
Temporal/1-Fix/4	38	3.60
Temporal/2-Fix/3	24	2.27
Fix/1-Temporal/1-Fix/3	19	1.80
Fix/2-Temporal/3	12	1.14
Aturat/1-Temporal/4	10	0.95
Fix/1-Temporal/4	10	0.95
Inactiu/5	10	0.95

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

La informació del quadre 3 es pot presentar de manera gràfica mitjançant la instrucció `seqfplot()` que permet veure ràpidament quines són les trajectòries més freqüents, que surten representades amb una àrea major o menor en funció de la seva freqüència (figura 3).

```
seqfplot(dades.seq, border=NA, withlegend=FALSE, title = "Fre-  (viii)
qüències")
```

Figura 3. Representació gràfica del 67,8% de les trajectòries laborals de les persones entrevistades al Panel de Desigualtats (PaD)



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

TraMineR ens ofereix una manera alternativa d'explorar quina és la seqüència més freqüent (moda), en aquest cas, la persona treballadora fixa en totes les rondes, de manera global i segons una variable explicativa d'interès:

```
seqmodst(dades.seq)                                    (ix)

by(dades.seq, dades$sexe, seqmodst)                    (x)
```

Els resultats d'executar aquests dos comandaments es poden trobar als quadres 4 i 5 respectivament. Tot i que la seqüència d'estat laboral "fix" és la més freqüent en ambdós sexes, se n'observa una major freqüència en els homes.

Quadre 4. "Moda" de les seqüències laborals

```
  [Modal state sequence]
  Sequence
1 Fix-Fix-Fix-Fix-Fix

  [State frequencies]
      status.1 status.2 status.3 status.4 status.5
Freq.     0.52     0.54     0.57     0.56     0.54
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Quadre 5. "Moda" de les seqüències laborals segons el sexe

```
dades$sexe: Home
  [Modal state sequence]
  Sequence
1 Fix-Fix-Fix-Fix-Fix

  [State frequencies]
      status.1 status.2 status.3 status.4 status.5
Freq.     0.64     0.67     0.69     0.67     0.65
------------------------------------------------------------
dades$sexe: Dona
  [Modal state sequence]
  Sequence
1 Fix-Fix-Fix-Fix-Fix

  [State frequencies]
      status.1 status.2 status.3 status.4 status.5
Freq.     0.43     0.44     0.47     0.47     0.45
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Les dades dels quadres 4 i 5 també es podrien representar de manera gràfica, `global (x)` i segons el `sexe (xi)`, mitjançant els comandaments següents:

```
seqmsplot(dades.seq, border=NA)                                    (xi)

seqmsplot(dades.seq, group=dades$sexe, border=NA)                  (xii)
```

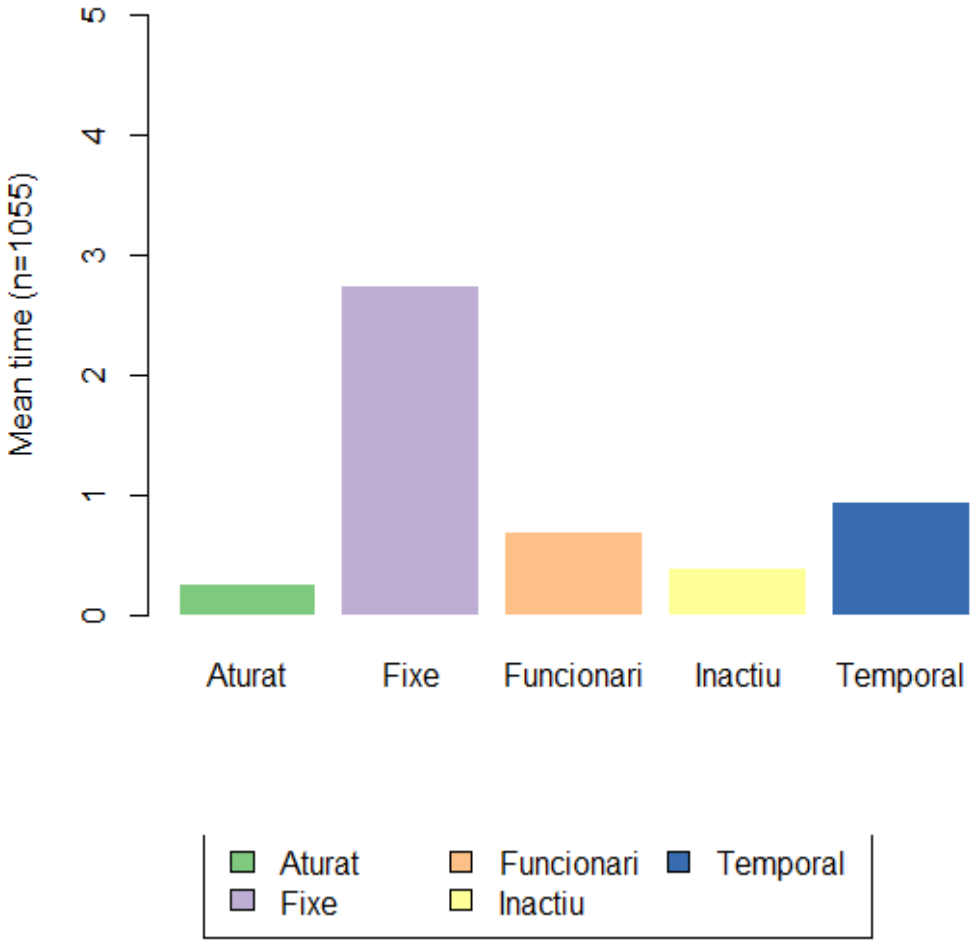
A més de representar les trajectòries individuals de cada persona, el paquet TraMineR també ens ofereix la possibilitat d'obtenir mesures resum, com per exemple, el temps mitjà en cadascun dels estatus laborals:

```
seqmtplot(dades.seq, border=NA)                                    (xiii)
```

Així, per exemple, observem un major temps mitjà (en anys) en la situació laboral "fix", seguit de "temporal" (figura 4). També podem obtenir aquest resultat segons les variables explicatives categòriques d'interès:

```
                                                                   (xiv)
seqmtplot(dades.seq, group=dades$sexe, border=NA)
```

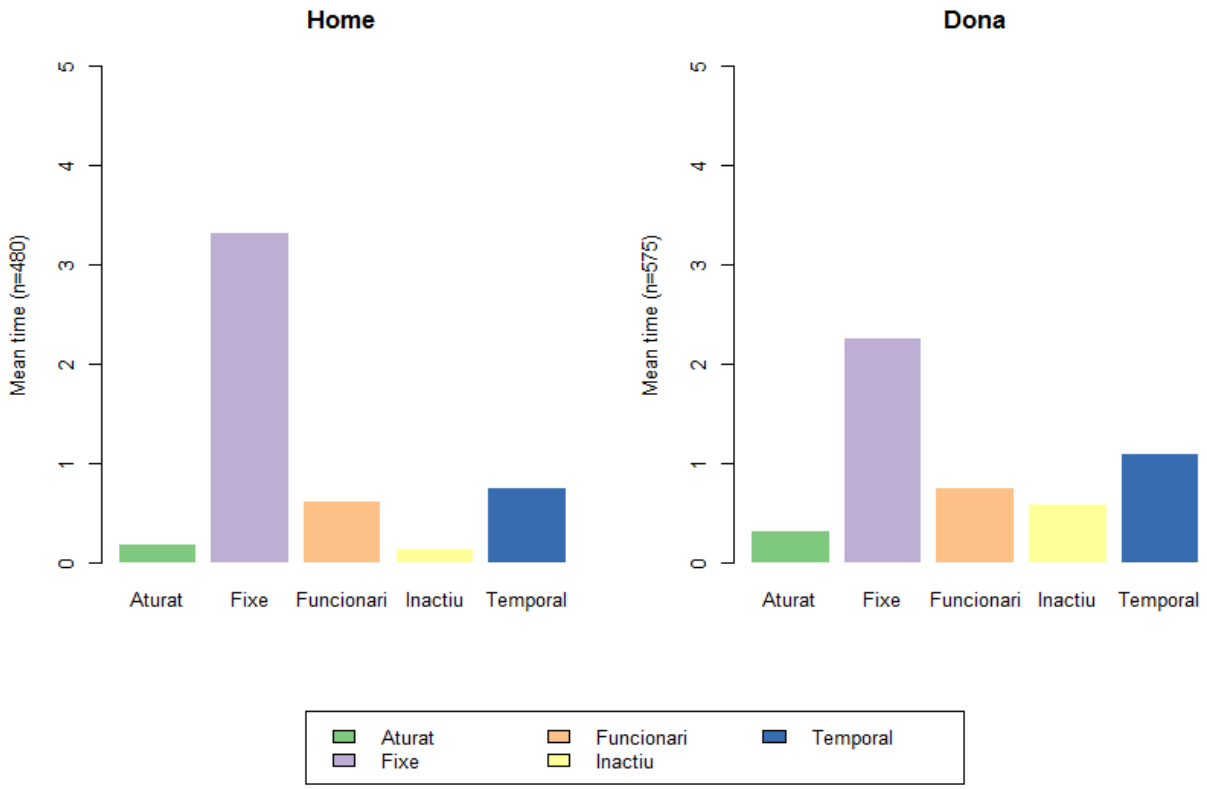
Figura 4. Temps mitjà de cada persona en cada situació laboral



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Així, a la figura 5 podem observar que, tot i que la distribució del temps mitjà en cada estatus laboral és similar entre els homes i les dones, cal destacar entre les dones una menor permanència en la situació laboral "fix", mentre que el temps mitja en les situacions d'inactivitat, temporalitat i atur és major.

Figura 5. Temps mitjà de cada persona en cada situació laboral segons el sexe



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

El temps mitjà segons el sexe es pot obtenir numèricament amb el comandament (quadre 6):

```
by(dades.seq, dades$sexe, seqmeant)                    (xv)
```

Quadre 6. Temps mitjà de cada persona en cada situació laboral segons el sexe

```
dades$sexe: Home
             Mean
Aturat     0.18
Fix       3.32
Funcionari 0.62
Inactiu    0.13
Temporal   0.75
------------------------------------------------------------
dades$sexe: Dona
             Mean
Aturat     0.31
Fix       2.26
Funcionari 0.74
Inactiu    0.59
Temporal   1.09
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Per últim, també podem explorar numèricament (comandament `xvi`) (quadre 7) i gràficament (comandament `xvii`) (figura 6) quina ha estat la situació laboral més freqüent en cadascuna de les rondes:

```
seqstatd(dades.seq)                                    (xvi)

seqdplot(dades.seq, border=NA)                         (xvii)
```

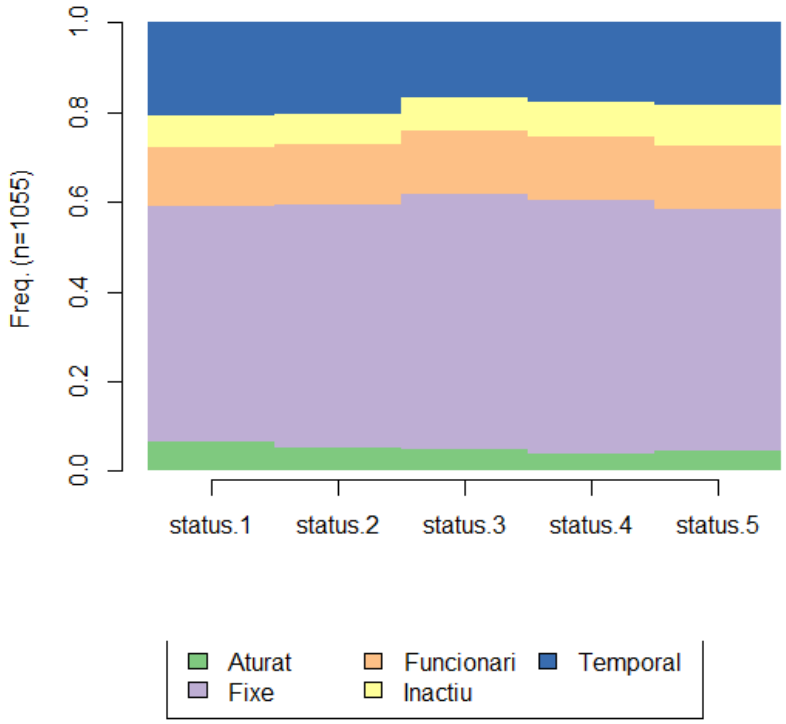
Al quadre 7 veiem com la situació laboral "fix" és la més freqüent en totes les rondes, seguida de la situació "temporal" i la situació de temporalitat. Cal destacar també l'estabilitat en la distribució de freqüències de les diferents situacions laborals en totes les rondes.

Quadre 7. Situació laboral més freqüent a cada ronda

```
                    [State frequencies]
           status.1 status.2 status.3 status.4 status.5
Aturat        0.066    0.052    0.047    0.040    0.045
Fix           0.524    0.543    0.570    0.564    0.539
Funcionari    0.132    0.132    0.140    0.140    0.141
Inactiu       0.071    0.069    0.076    0.078    0.089
Temporal      0.207    0.204    0.167    0.178    0.185


           status.1 status.2 status.3 status.4 status.5
N              1055     1055     1055     1055     1055

                     [Entropy index]
           status.1 status.2 status.3 status.4 status.5
H              0.81     0.78     0.77     0.77     0.79
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Figura 6. Situació laboral més freqüent a cada ronda

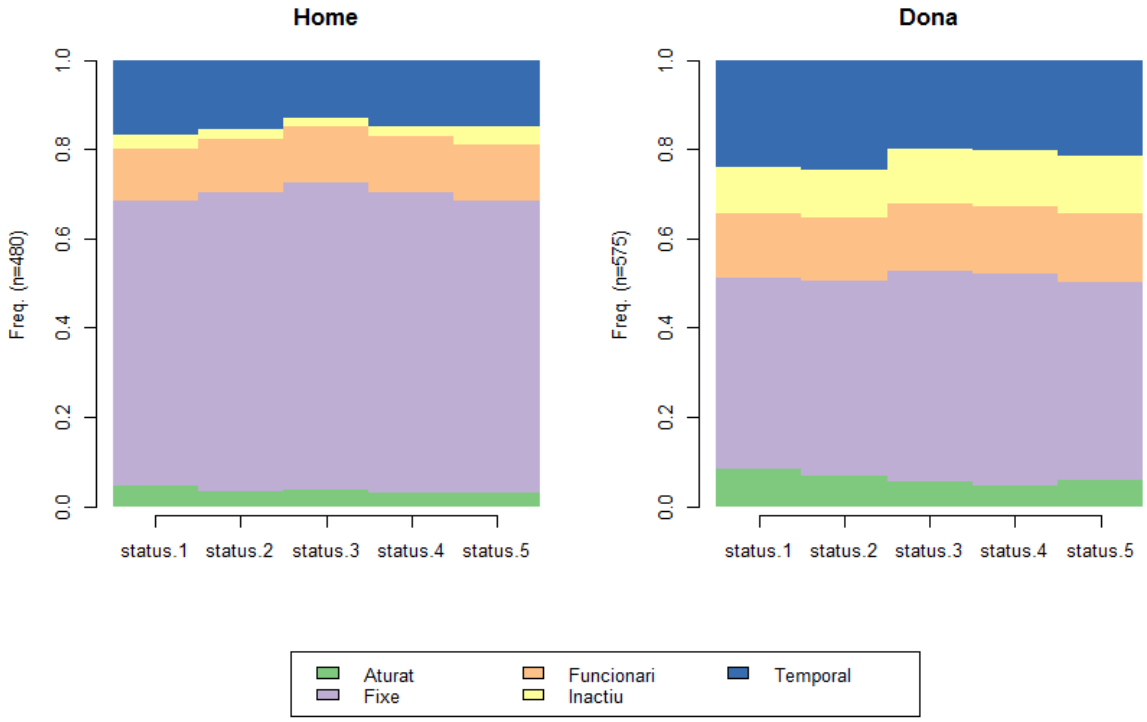


Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Com hem vist amb altres funcions del TraMineR, podem analitzar la distribució dels diferents estats de situació laboral a cada ronda segons les variables explicatives:

```
seqdplot(dades.seq, group=dades$sexe, border =NA)                    (xviii)
```

Figura 7. Situació laboral més freqüent a cada ronda segons el sexe



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

A la figura 7 observem com els homes i les dones tenen distribucions similars, no obstant això, cal destacar una major freqüència d'inactivitat, atur i temporalitat en les dones en les cinc rondes d'estudi.

Al quadre 7, a més de la distribució de l'estatus o situació laboral en cada ronda, el comandament `(xvi)` ens permet obtenir l'índex d'entropia o entropia de Shanon per a cada estatus laboral. L'entropia de Shanon es defineix com:

$$h(p_1,\ldots,p_a) = -\sum_{i=1}^{a} p_i \log(p_i)$$

Essent a = el nombre d'estats i p_i la proporció de casos a l'estat i (i=1,..., 5) a cada ronda.

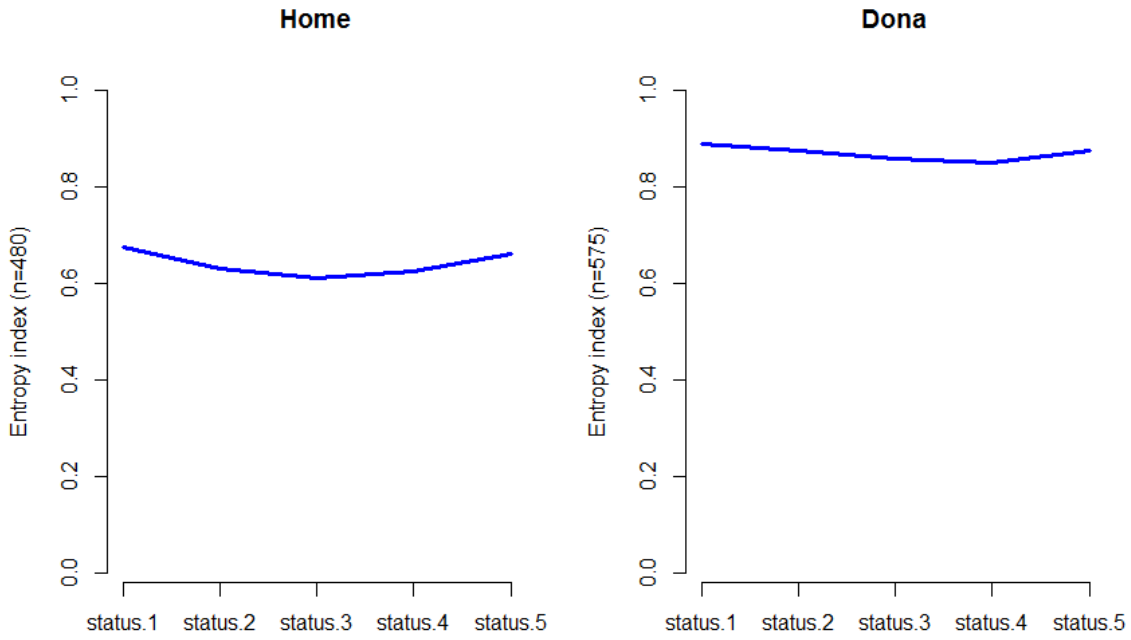
L'entropia prendrà el valor 0 quan totes les observacions tenen el mateix estatus i el valor màxim 1 quan hi hagi la mateixa proporció d'observacions a tots els estats. Així, l'entropia es podria interpretar com una mesura de la diversitat en la situació laboral a cada ronda.

En l'exemple que estem estudiant, obtenim entropies properes a 0,8 en les cinc rondes; podríem considerar doncs que hi ha una diversitat elevada en els estats observats a cada ronda. Les entropies de cada ronda es poden mostrar de manera gràfica i distingint-les segons les variables explicatives d'interès amb el comandament `(xix)`. Aquest tipus de gràfics són especialment interessants per analitzar l'evolució temporal de l'entropia (figura 8).

```
SeqHplot(dades.seq, group=dades$sexe)                    (xix)
```

A la figura 8 veiem com la variabilitat d'estats a cada ronda és major entre les dones que entre els homes. D'altra banda, també podem observar com l'entropia disminueix fins a la quarta ronda en ambdós sexes, i a partir d'aquesta sembla que torna a augmentar. És interessant comparar aquesta informació amb l'obtinguda a partir de la figura 7, ja que, en efecte, la freqüència d'estats en cada ronda en les dones tendeix a ser més similar que en els homes (exceptuant l'estat "fix", que és el més freqüent en ambdós sexes).

Figura 8. Entropia de cada ronda segons el sexe



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Construcció d'agrupacions d'individus segons les seves trajectòries laborals

Per a la construcció d'agrupacions d'individus d'acord amb les seves trajectòries laborals caldrà definir una forma de calcular la distància entre les diferents situacions laborals. En aquest sentit, en primer lloc, caldrà definir la **taxa o probabilitat de transició** com la probabilitat de passar d'un estat s_i a un estat s_j, on els estats serien les diferents possibilitats de situació laboral (inactiu, aturat, temporal, fix i funcionari). Matemàticament, questes probabilitats o taxes de transició $p(s_i, s_j)$ s'obtenen de l'expressió:

$$p(s_j / s_i) = \frac{\sum_{t=1}^{L-1} n_{t,t+1}(s_i, s_j)}{\sum_{t=1}^{L-1} n_t(s_i)}$$

Essent:

L: la longitud de la seqüència,
$n_t(s_i)$: el nombre de seqüències que no finalitzen en el moment *t* amb l'estat s_i, en el moment *t* i $n_{t,t+1}(s_i,s_j)$ el nombre de seqüències amb l'estat s_i al moment *t* i l'estat s_j al moment *t*+1.

El TraMineR proporciona les taxes de transició amb una matriu on la fila *i* proporciona la distribució de la transició des de l'estat s_i en el moment *t* a la resta d'estats en el moment *t*+1 (en les columnes), de manera que cada fila de la matriu suma 1. Així, les taxes de transició proporcionen informació sobre els canvis d'estats més freqüents a les dades i també ens permeten veure quina és l'estabilitat de cada estat, amb els valors a la diagonal, que ens informen de la taxa de transició d'un estat s_i al moment *t*, al mateix estat s_i al moment *j*.

Vegem-ho en el nostre exemple, mitjançant el comandament `seqtrate()`:

```
round(seqtrate(dades.seq),2)                                    (xx)
```

La funció `round()` ens permet arrodonir les probabilitats de transició a dos decimals.

Quadre 8. Probabilitats de transició d'una situació laboral a una altra

```
[>] computing transition rates for states Aturat/Fix/Funcionari/Inactiu/Temporal
...

                [-> Aturat] [-> Fix] [-> Funcionari] [-> Inactiu] [-> Temporal]
[Aturat ->]            0.27      0.15            0.01         0.22          0.35
[Fix ->]               0.02      0.92            0.01         0.01          0.04
[Funcionari ->]        0.00      0.02            0.96         0.00          0.02
[Inactiu ->]           0.14      0.08            0.01         0.60          0.18
[Temporal ->]          0.07      0.17            0.02         0.08          0.67
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Al quadre 8 tenim la matriu de transicions. En efecte, observem que les files sumen 1. D'altra banda, veiem que les probabilitats majors es troben sobretot a la diagonal, en les situacions laborals "fix" (0,92), "funcionari" (0,96), "inactiu" (0,60) i "temporal" (0,67). Aquests serien, per tant, les situacions laborals més estables en la nostra mostra, mentre que l'estat d'"aturat" seria el més inestable, i tendiria a moure's cap a l'estatus de "temporal" (0,35), "inactiu" (0,22) i "fix", en menor mesura (0,15). Recordem que aquests resultats són similars als obtinguts en l'estudi descriptiu previ.

Observem també la coherència de la matriu de transicions, atès que, per exemple, la probabilitat de passar de "funcionari" o "fix" a qualsevol altre estat és molt petita.

A continuació veurem com podem realitzar una agrupació o clúster d'individus, a partir del mètode *Optimal Matching* o de correspondència òptima. El mètode de l'*Optimal Matching* ens permet definir una forma de calcular distàncies entre trajectòries, i s'ha presentat com una forma alternativa al càlcul de distàncies estàndard utilitzat en els mètodes factorials propis de l'anàlisi multivariant (Massoni *et al.*, 2009).

El TraMineR permet utilitzar altres mètodes de càlcul de distàncies entre trajectòries laborals. Per a més informació es pot consultar el capítol 9 de Gabadinho, Ritschard,

Studer i Müller (2011).

Les distàncies d'*Optimal Matching* (OM), també conegudes com a *edit distances* o *Levenshtein distances*, es van començar a utilitzar en les ciències biològiques i van ser introduïdes al camp de les ciències socials per Abbott i Forrest (1986). Aquest mètode es basa en la definició de costos, entesos com el nombre d'operacions que calen per transformar una seqüència d'estats a_i a una altra a_j. Aquestes operacions inclouen la inserció, substitució o eliminació d'estats. En el nostre cas, atès que tenim seqüències completes per a tots els individus, és a dir, tots els individus tenen seqüències de cinc possibles estats a les cinc rondes, només caldrà calcular els costos de substitució.
Els *costos de substitució* es defineixen com:

$$w(s_i,s_j) = 2\text{-}p(s_i|s_j)\text{-}p(s_j|s_i)$$

essent s_i i s_j dos possibles estats i $p(s_i|s_j)$ les probabilitats de transició entre els dos estats. Així, com més petita sigui la probabilitat de transició per passar d'un estat s_i a un altre s_j o viceversa, més alt serà el cost de passar de l'estat s_i a l'estat s_j, o viceversa. Dit d'una altra manera, com menys transicions s'observen entre dos estats laborals, els estats es consideraran més diferents, i el cost de substitució serà major.

Per a més informació sobre la definició dels costos es pot consultar Gauthier Cédric Notredame (2009).

Per a calcular la matriu de costos de substitució amb TraMineR podem fer servir el comandament `seqsubm()`, considerant el mètode TRATE, corresponent a les probabilitats de transició (quadre 9).

```
round(seqsubm(dades.seq,method="TRATE"),2)                    (xxi)
```

Quadre 9. Costos de substitució d'un estat $_i$ a un estat $_j$

```
 [>] creating substitution-cost matrix using transition rates ...
[>]     computing     transition     rates     for     states     Atu-
rat/Fix/Funcionari/Inactiu/Temporal ...
             Aturat-> Fix-> Funcionari-> Inactiu-> Temporal->
Aturat->         0.00   1.83         1.99      1.65       1.58
Fix->           1.83   0.00         1.97      1.91       1.79
Funcionari->     1.99   1.97         0.00      1.99       1.97
Inactiu->        1.65   1.91         1.99      0.00       1.74
Temporal->       1.58   1.79         1.97      1.74       0.00
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Al quadre 9, observem com el mínim cost, 0, el trobem en la substitució de cada estat pel mateix i el màxim és menor que 2; atès que 2 seria el valor del cost que obtindríem en una transició no observada en la nostra mostra. D'acord amb els resultats obtinguts a la matriu de transició (quadre 8), observem els costos menors quan substituïm l'estat "aturat" per "temporal" o "inactiu". Mentre que el cost és major quan passem d'"aturat a "funcionari" o bé de "funcionari" a "inactiu".

Ja tenim, doncs, tots els elements bàsics per a poder elaborar una agrupació de seqüències. L'agrupació de seqüències és un mètode exploratori d'anàlisi de dades amb l'objectiu de trobar grups homogenis d'individus (Kaufman i Rousseeuw, 2005). En aquest cas, considerarem les distàncies OM entre estats per determinar aquestes agru-

pacions.

En primer lloc, caldrà definir la matriu de distàncies entre seqüències d'estats, d'acord amb la definició de costos de substitució per passar d'un estat a un altre:

```
dades.om<-seqdist(dades.seq, method="OM", sm ="TRATE", (xxii)
full.matrix=FALSE)
```

Quan s'executa el comandament (xxii), obtenim la informació del quadre 10, amb una descripció de les dades i els mètodes utilitzats.

Quadre 10. Confirmació de la construcció de la matriu de costos de substitució

```
[>] creating substitution-cost matrix using transition rates ...
 [>] computing transition rates for states Atu-
rat/Fix/Funcionari/Inactiu/Temporal ...
 [>] 1055 sequences with 5 distinct events/states
 [>] 197 distinct sequences
 [>] min/max sequence length: 5/5
 [>] computing distances using OM metric
 [>] total time: 0.17 secs
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

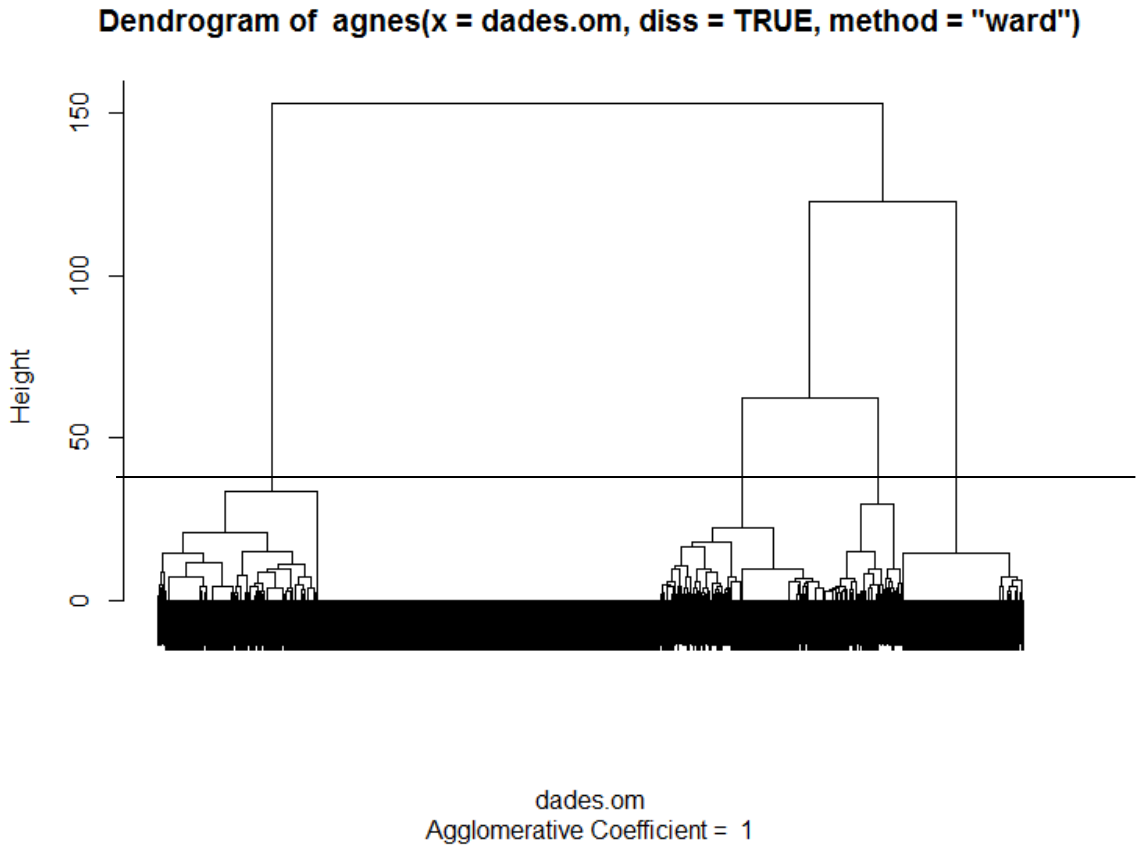
A continuació, podem procedir a realitzar una anàlisi de clúster jeràrquic, amb la matriu de distàncies obtingudes. Per a poder fer aquesta anàlisi podem utilitzar la funció agnes() de la llibreria cluster, considerant, per exemple, el mètode de Ward. El resultat de l'anàlisi el podem guardar en un objecte de R, per a facilitar les anàlisis posteriors. En aquest cas, l'anomenarem clusterdades (xxiii).

```
library(cluster) (xxiii)
clusterdades<-agnes(dades.om, diss=TRUE, method="ward")
```

Per a decidir el nombre òptim d'agrupacions de seqüències representarem el dendrograma corresponent:

```
plot(clusterdades, which.plot=2, labels=FALSE) (xxiv)
```

Figura 9. Dendrograma per a l'agrupació de les trajectòries laborals. Panel de Desigualtats (PaD). Catalunya, 2002-2006



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

A la figura 9 observem que un nombre òptim d'agrupacions podria ser quatre. Establim quina serà l'agrupació òptima amb la representació del dendograma. Procedim, per tant, a definir aquests clústers d'individus:

```
dades.c14<-cutree(clusterdades, k=4)                    (xxiv)
```

Per acabar, podem convertir la variable de classificació obtinguda en un factor:

```
c14.lab<-factor(dades.c14, labels=paste("Cluster",1:4))    (xxv)
```

D'aquesta manera, veiem que les 1.055 persones entrevistades al PaD en les cinc rondes, entre el 2002 i el 2006, se'ns han classificat en quatre grups.

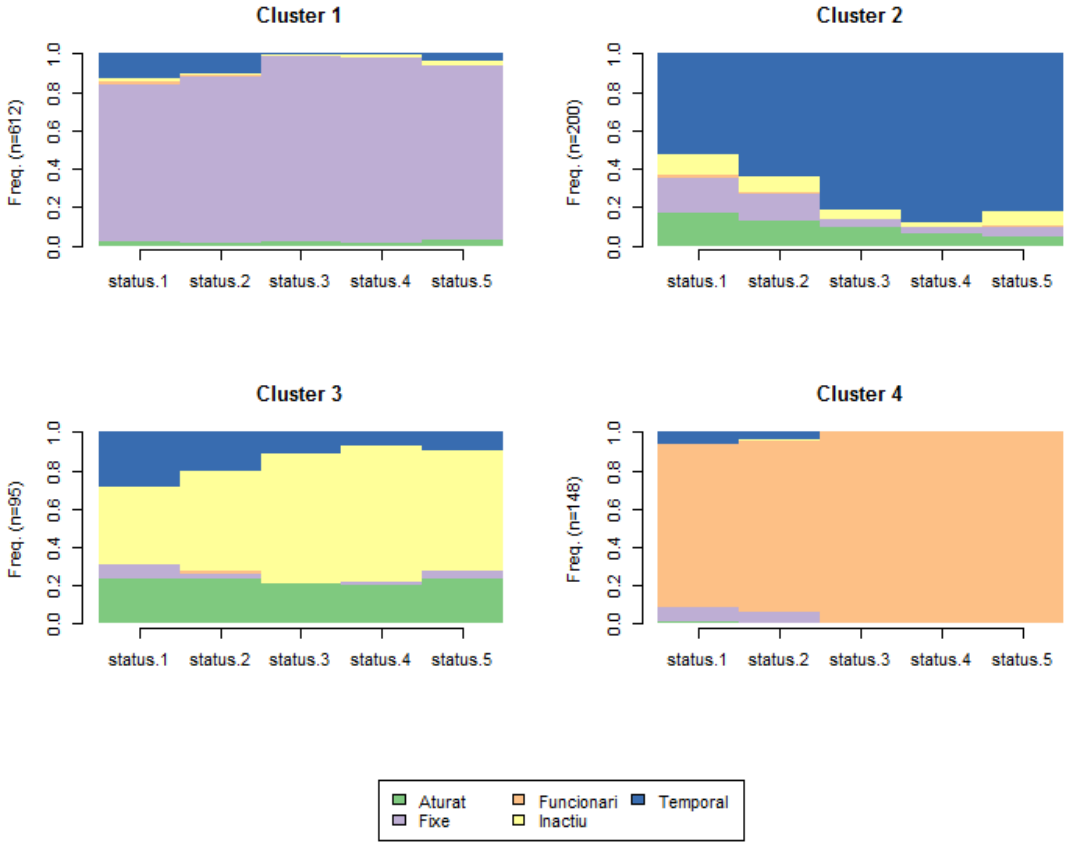
Finalment, amb la instrucció `seqdplot()` podem veure quin és el patró d'aquests clústers d'acord amb la distribució transversal dels diferents estats laborals:

```
seqdplot(dades.seq,group=c14.lab,border=NA)              (xxvi)
```

A la figura 10 es poden distingir quatre grups diferenciats, sent el primer grup, o clús-

ter 1, el més nombrós, amb 612 de les 1.055 persones entrevistades. Aquest grup es correspon essencialment amb les persones amb treball fix durant tot el període d'estudi. També s'observa un altre clúster força estable, el clúster 4, que es correspon amb les persones funcionàries. D'altra banda, s'observa un grup de persones majoritàriament inactives durant les cinc rondes, o bé en situacions d'atur o temporalitat (clúster 3) i, per últim, un altre grup (el clúster 2), essencialment en situació de temporalitat o amb tendència a la temporalitat. Aquests dos darrers grups serien els que podríem considerar, per tant, els més precaris o inestables quant a la seva trajectòria laboral.

Figura 10. Clústers de trajectòries laborals. Panel de Desigualtats (PaD). Catalunya, 2002-2006



Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

D'acord amb la descripció obtinguda a partir de la figura 10, podem reetiquetar els clústers de manera més intuïtiva:

```
c14.lab<-factor(dades.c14,labels=c("C1-Fixos","C2-Temporals","C3-Inactius","C4-Funcionaris")
```
(xxvii)

Caracterització del les agrupacions de trajectòries laborals en funció de variables explicatives

Finalment, podem arribar a caracteritzar cadascun dels quatre clústers en funció de les variables explicatives que hem seleccionat per a aquest estudi de cas (sexe, edat, nivell d'estudis a l'inici de l'estudi i ingressos mensuals al final de l'estudi). En primer lloc, farem una anàlisi descriptiva bivariada, d'acord amb la naturalesa de les variables expli-

catives (comandament xxviii) (quadre 11):

```
with(dades, table(sexe, c14.lab))                                    (xxviii)

with(dades, table(estudis, c14.lab))

with(dades, do.call(rbind, tapply(edatinici, c14.lab, functi-
on(x) c(Mitjana = mean(x), DT = sd(x)))))

with(dades, do.call(rbind, tapply(ingressosfi, c14.lab, func-
tion(x) c(Mitjana = mean(x), DT = sd(x)))))
```

Quadre 11. Anàlisi descriptiva de la distribució de les variables explicatives en les quatre agrupacions o clústers

```
Sexe:
sexe   C1-Fixos C2-Temporals C3-Inactius C4-Funcionaris
  Home      332           71          16             61
  Dona      280          129          79             87

Estudis:
                                          c14.lab
estudis                                        C1-Fixos C2-Temporals C3-
Inactius
  'Primària o menys'                                105            36
25
  'Primer cicle de secundària'                      112            39
28
  'Segon cicle de secundària, no tècnic'             82            27
12
  'Segon cicle de secundària, tècnic'                93            33
15
  'Educació postsecundària, no terciària'            80            15
6
  'Estudis superiors'                               140            50
9
                                         c14.lab
estudis                                   C4-Funcionaris
  'Primària o menys'                                   3
  'Primer cicle de secundària'                         3
  'Segon cicle de secundària, no tècnic'              18
  'Segon cicle de secundària, tècnic'                  5
  'Educació postsecundària, no terciària'              7
  'Estudis superiors'                                112

Edat:
               Mitjana        DT
C1-Fixos       38.62745  9.826403
C2-Temporals   34.18500  9.935004
C3-Inactius    37.50526 12.302805
C4-Funcionaris 42.86486  7.464791

Ingressos:
                 Mitjana       DT
C1-Fixos       1437.6355 781.6255
C2-Temporals   1132.8985 655.2959
C3-Inactius     442.8964 557.2551
```

```
C4-Funcionaris 1960.7885 823.9403
```

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

Al quadre 11 observem que en les agrupacions de trajectòries més favorables (fixos i funcionaris) destaquen sobretot els homes d'edat mitjana superior i amb estudis superiors. D'altra banda, són també aquests grups els que reben salaris mensuals menors. En canvi, les agrupacions de trajectòries més precàries es caracteritzen per incloure les dones, els més joves, amb menors ingressos mensuals i menor nivell d'estudis.

Vegem finalment com podríem quantificar les diferències observades entre els clústers i les variables explicatives, mitjançant un model logístic multinomial, considerant el clúster C3, d'inactivitat, com a referència:

```
c14.lab3<- relevel(c14.lab, ref = "C3-Inactius")                    (ixxx)
```

A continuació ajustem el model multinomial, carregant prèviament la llibreria `nnet`.

```
library(nnet)                                                        (xxx)
model        <-        multinom(c14.lab3        ~        se-
xe+estudis+edatinici+ingressosfi, data = dades)
```

I amb els comandaments següents obtenim les oportunitats relatives (*odds ratios*) i els valors de *p* corresponents:

```
or<-exp(coef(model))                                                 (xxxi)
round(or,3)
z<-summary(model)$coefficients/summary(model)$standard.errors
round(z,3)
p<-(1-pnorm(abs(z),0,1))*2
round(p,3)
```

Els resultats es resumeixen a la taula 2. Es confirmen els resultats obtinguts a l'anàlisi descriptiva preliminar. Concretament, observem que l'*odds* d'estar al grup de persones amb treball fix vs. les persones inactives és major entre els homes, de major edat, amb estudis secundaris tècnics o superiors i amb majors ingressos. Aquesta distribució és similar a l'observada quan es compara el grup de persones funcionàries vs. les persones inactives, on tenir estudis per sobre dels primaris, tenir major edat o ingressos suposa una major *odds* de pertànyer al grup de persones funcionàries. En aquest grup, no s'observen, no obstant això, diferències segons el sexe.

Per acabar, quan analitzem el grup de persones amb treball temporal vs. persones inacties, tampoc no s'observen diferències segons el sexe. Respecte a la resta de variables explicatives, l'*odds* de tenir majors ingressos és major entre el grup de temporals, mentre que, per l'edat, l'*odds* és major en el grup d'inactius. Quant al nivell d'estudis, s'observa una major *odds* d'estar inactiu vs. temporal amb estudis inferiors (secundària i no tènics), o bé amb estudis postsecundaris.

L'aproximació a la caracterització de les quatre agrupacions obtingudes en aquest exemple és similar a un estudi previ realitzat també amb les dades del PaD en el període 2002-2006, amb una manera alternativa de resumir i agrupar les trajectòries laborals (Verd i López-Andreu, 2012).

Taula 2. Models de regressió logística multinomial corresponents als quatre clústers de trajectòries laborals (Ref. inactius, homes, sense estudis o estudis primaris). Panel de Desigualtats (PaD). Catalunya, 2002-2006

	Fixos			Temporals			Funcionaris		
	OR	Z	P-VALOR	OR	Z	P-VALOR	OR	Z	P-VALOR
Constant	0,187*	-11018,000	<0,001	0.6384*	-3,884	<0,001	0,000*	-225,062	<0,001
Sexe									
Dones	0,715*	-3,054	0,002	1,218	1,437	0,151	1,085	0,619	0,536
Edat	1,015*	2,147	0,032	0,967*	-4,521	<0,001	1,090*	10,450	<0,001
Estudis									
1r secundària	0,874	-1,215	0,224	0,630*	-4,924	<0,001	1,356*	9,580	<0,001
2n secundària, no tècnic	0,790	-1,769	0,077	0,631*	-2,676	0,008	8,177*	12,724	<0,001
2n secundària, tècnic	1,449*	3,103	0,002	1,147	1,286	0,199	3,412*	17,929	<0,001
Postsecundària, no terciària	1,892*	4,581	<0,001	0,756*	-2,315	0,021	8,707*	23,779	<0,001
Superiors	1,398*	3,164	0,002	1,135	0,943	0,361	54,695*	30,843	<0,001
Ingressos	1,004*	12,103	<0,001	1,003*	10,195	<0,001	1,004*	12,301	<0,001

Font: Panel de Desigualtats Socials a Catalunya, 2002-2006.

3.2. Models de coeficients aleatoris: factors predictors dels ingressos i evolució al llarg del període 2003-2009

En aquest exemple s'ajusten models de coeficients aleatoris per determinar les variables significativament associades al nivell d'ingressos en una submostra de persones ocupades i estudiar-ne l'evolució al llarg del període 2003-2009.

El programari utilitzat ha estat Stata v.12.1 Special Edition.

3.2.1. Plantejament de l'estudi

Com hem remarcat anteriorment, l'elecció del model estadístic a utilitzar ve determinat principalment per les preguntes que es planteja la nostra recerca. En aquest exemple ens preguntem en primer lloc pels factors que determinen el nivell d'ingressos real en una submostra de persones ocupades (models A) i, en segon lloc, ens centrem en l'estudi de l'evolució dels ingressos al llarg del període d'observació i la seva variabilitat al nivell individual (model B1).

Cal observar que la primera qüestió podria també respondre's a partir de dades transversals. Tanmateix, en un plantejament longitudinal podem fer que cada subjecte actuï com el seu propi control. D'aquesta manera, es controla l'efecte de característiques individuals no mesurades (per exemple el coeficient intel□lectual, la capacitat de treball o la motivació) que poden afectar els ingressos i les conclusions obtingudes tenen un major nivell de validesa.

En contrapartida, i des del punt de vista estadístic, les dades longitudinals presenten el problema de l'autocorrelació entre les mesures de cada individu, circumstància aquesta que viola un dels supòsits fonamentals de la regressió. Els models A1, A2 i A3 adrecen aquest problema de manera diferent, mitjançant mínims quadrats generalitzats, *intercept* aleatori i efectes fixos.

Quant a la segona qüestió, l'evolució dels ingressos al llarg del període, només la disponibilitat de dades longitudinals permet estudiar el canvi intraindividual i adreçar la pregunta sobre diferències interindividuals al canvi intraindividual dels ingressos. El model B1 utilitza un model de pendent aleatori per donar resposta a aquesta pregunta.

Selecció de la mostra

Les dades d'aquest exemple (SELINGRE.dta) corresponen a una submostra d'individus del PaD corresponent a persones d'entre 16 i 64 anys que s'han mantingut permanentment ocupades al llarg del seu temps efectiu d'observació (és a dir entre un i set anys) a partir del 2003 i fins al 2009. S'han eliminat també els individus amb valor perdut en les covariables incloses en el model. La mostra d'anàlisi consta finalment de n = 1.316 individus i 6.317 registres.

Cal tenir en compte per tant que no es tracta d'una mostra representativa del conjunt de la població activa, sinó d'un subconjunt comparativament privilegiat que no s'ha vist afectat per l'atur ni situacions d'inactivitat. La mediana d'ingressos després de la selecció és per exemple de 1.400 € (rang interquartílic 900 €), mentre que en el conjunt de la població activa (valors perduts i mestresses de casa exclosos) era de 1.125 € (rang interquartílic 929 €). Les estimacions es trobaran per tant profundament afectades per un biaix de selecció.

Variables en estudi

Les variables referents al salari del PaD contenen un nombre elevat de valors perduts la proporció dels quals s'accentua en un plantejament longitudinal. Per aquesta raó vam decidir utilitzar com a proxy del salari la variable codi I140002EG1 "ingressos nets propis (any anterior) imputada (euros/mes)", disponible per a les onades 2003-2009. Es trata d'una variable imputada construida pel Servei d'Estadística de la Universitat Autònoma de Barcelona i "socialitzada" pel PaD. Existeix documentació tècnica referent a la construcció d'aquesta variable (Servei d'Estadística UAB, 2012).

Les variables explicatives que considerarem en el nostre model són les següents: sexe, nivell educatiu, anys treballats, antiguitat en la feina actual (*tenure*), treball a temps complet o parcial i número d'onada (la corresponent al 2003 es considera onada zero). A la taula 3 es resumeixen les variables d'estudi.

Taula 3. Variables d'estudi

Tipus	Nom	Descripció	Valors
Dependent	**ingressos**	Ingressos mensuals nets any anterior	Rang: 99 a 12875
Independents	**onada**	Onada	Rang: 0 a 6
	sexe	Sexe de la persona entrevistada	0: Home 1: Dona
	estudis_1	Nivell màxim d'estudis assolit fins al moment de l'entrevista	1: Primària o menys 2: Primer cicle de secundària 3: Segon cicle de secundària, tècnic 4: Segon cicle de secundària, no tècnic 5: Educació postsecundària, no terciària 6: Estudis superiors
	anys_treb	Anys transcorreguts des que es va comencar a treballar remuneradament	Rang: 1 a 54
	tenure3	Anys que fa que es treballa a l'empresa (des de la darrera interrupció)	Rang: 0 a 49
	intensitat	Treball a temps complet o parcial	0: Parcial 1: Complet

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Cal destacar que entre les covariables seleccionades apareixen diverses mètriques del temps: onada, anys treballats i *tenure*. Qualsevol d'elles es pot fer servir com a mesura del canvi. Si ens centrem en la primera tendirem a veure l'efecte del període sobre el salari; la segona ens condueix a considerar l'efecte de la permanència en el mercat de treball, mentre que la tercera ens mesura l'efecte de la permanència a l'empresa. Allò que hem descartat és l'existència d'un efecte generacional, és a dir, els possibles efectes sobre l'ingrés actual de pertànyer a una determinada cohort d'entrada al mercat de treball. Aquesta solució suposa acceptar la constància dels efectes de la cohort d'entrada al mercat de treball sobre els ingressos actuals.

3.2.2. Caracterització dels models estadístics

3.2.2.1. Questió A: factors determinants del nivell d'ingressos

Model A1. Regressió per mínims quadrats ordinaris (OLS)

La regressió per mínims quadrats ordinaris (Ordinary Least Squares, OLS) dóna estimacions consistents dels paràmetres poblacionals (és a dir, el seu valor esperat és el del paràmetre) si es compleix que els residus de la regressió són independents i idènticament distribuïts.

Si aquest supòsit no es compleix, la matriu de variàncies-covariàncies (VCE) no es pot estimar adequadament, cosa que es traduirà en un interval de confiança dels estimadors incorrecte. En el nostre cas, atès que amb dades longitudinals és esperable que les mesures corresponents a un mateix subjecte estiguin correlacionades, els residus de la regressió també ho estaran, i el supòsit que es distribueixen independentment no es complirà.

En aquesta situació, una possibilitat és considerar que les estimacions puntuals obtingudes mitjançant OLS són consistents, però que cal corregir l'estimació de la seva variança (Baum, 2006, p. 133).

El Model A1 és una regressió per mínims quadrats ordinaris en què els errors de les estimacions es calculen mitjançant un estimador robust front a l'existència de correlació en les dades. No proporcionem més detalls ja que l'estimació robusta no constitueix el focus d'aquest exemple.

Model A2. Intercept aleatori (random intercept model)

El model A2 adreça igualment el problema de l'estimació dels coeficients de regressió amb dades correlacionades, però el resol de forma diferent a l'anterior.

En el marc de la recerca no experimental (anàlisi de la covariació), sovint apareixen diferències considerables al nivell individual (heterogeneitat individual) no explicades per la part fixa del model.

Aquesta heterogeneïtat suggereix l'existència d'importants factors predictors no inclosos en el model, i que sovint no poden incloure's-hi a caus de la dificultat per a mesurar-los, o altres causes. L'heterogeneïtat individual és una font d'endogeneïtat, la qual causa al seu torn inconsistència en les estimacions i dificulta la interpretació de les mateixes en un sentit causal.

Quan s'apliquen a dades longitudinals, els models d'intercepts aleatoris (random intercept models) constitueixen una manera natural d'afrontar el problema de l'heterogeneïtat individual.[14]

En ells, se suposa que existeix en cada subjecte un factor específic i constant en cada subjecte que influeix en el valor de la variable dependent. L'acció d'aquest factor ocasiona que les mesures d'un mateix subjecte siguin similars (estiguin correlacionades) i, alhora, siguin diferents de les de la resta d'individus (provoca heterogeneïtat individual).

Com es tradueix aquest supòsit en termes de model estadístic? Tot model de regressió consta d'una part fixa (les variables predictives i els seus coeficients) i una part aleatòria (el terme d'error). De forma general, en els models d'efectes aleatoris es divideix el terme d'error en variabilitat residual pròpiament dita i un o més factors aleatoris els paràmetres dels quals poden ser estimats per separat.

[14] Els models d'efectes aleatoris són una possibilitat de tractar l'autocorrelació de les dades ocasionada per la seva pertinença a un mateix conglomerat. En el cas de dades longitudinals, l'individu fa el paper de conglomerat.

En els models d'intercept aleatori, s'introdueix un únic factor aleatori que afecta la constant o intercept del model de regressió (o, dit d'un altra manera, el valor inicial de l'individu en la variable dependent). Mentre en la regressió ordinària tant l'intercept com el pendent de la recta (o pla o hiperplà) de regressió se suposa que són comuns a tots els individus, en els models d'intercept aleatori es permet als individus diferir en el valor de l'intercept en una certa quantitat aleatòria respecte al valor estimat poblacional. El pendent del model, en canvi, és comú a tots els individus.

$$y_{it} = (\beta_0 + \zeta_{0i}) + x_{1it}\beta_1 + x_{2it}\beta_2 + \ldots + x_{kit}\beta_k + \varepsilon_{it}$$

Part fixa del model

β_0 representa l'intercept poblacional,

$x_{1it},\ldots,x_{kit}$ representen els vectors de valors respectius de les variables $x_1,\ldots,x_k$ des del moment t=0 fins al màxim de t.

$\beta_1,\ldots,\beta_k$ representa els coeficients poblacionals de les variables $x_1,\ldots,x_k$

Part aleatòria del model

ζ_{0i} representa la desviació de l'individu respecte a l'intercept poblacional β_0 (és a dir, una variabilitat interindividual aleatòria específica de cada subjecte i constant al llarg de tots els moments t del temps).

ε_{it} representa l'error residual (variabilitat intraindividual), específic de cada individu-ocasió. Dit d'una altra manera, és la variació de l'individu respecte a la seva veritable trajectòria de canvi (que aquí suposem lineal).

Aquest model fa a més les assumpcions següents:

1. Condicionades a les covariables, les ζ_{0i} es distribueixen normalment i la seva esperança és zero.
2. Condicionades a les covariables i al factor aleatori, les ε_{it} es distribueixen normalment i la seva esperança és zero.
3. ζ_i i ε_{it} no estan correlacionades.
4. Les ζ_{0i} no estan correlacionades entre unitats d'anàlisi.
5. Les ε_{it} no estan correlacionades entre si ni dintre de les unitats d'anàlisi.

La introducció de l'intercept aleatori permet augmentar la consistència i precisió de les estimacions en cas d'heterogeneïtat individual, així com obtenir una mesura de la magnitud d'aquesta. En el

nostre cas, el model permet tenir en compte diferències interindividuals en el nivell d'ingressos degudes a factors personals constants no considerats en l'equació de regressió.

3.2.2.2. Qüestió B: evolució intraindividual dels ingressos i diferències interindividuals en la mateixa

Model de coeficients aleatoris (Random Coefficient Model)

En els models d'efectes aleatoris[15], la introducció de factors aleatoris no té per què limitar-se a l'intercept.

En el model A1 d'intercept aleatori, suposem que els coeficients de les variables predictores són els mateixos per a tots els individus i idèntics als coeficients poblacionals. Aquesta condició tan restrictiva es pot relaxar, introduint un o més factors aleatoris que modulen el valor dels coeficients de les variables predictores al nivell individual. Això permet que la recta de regressió de cada individu (si estem ajustant una funció lineal, com és el cas més habitual) no tan sols difereixi en l'intercept, sinó en el seu pendent.

Un cas particular i particularment popular dels models de coeficients aleatoris es produeix quan és el coeficient de la variable "temps" allò que es permet que varïi aleatòriament entre els subjectes de la mostra. Aquests models són coneguts com a models de corbes de creixement (growth-curve models, latent-trajectory models o latent growth curve models) (Rabe-Hesketh i Skrondal, 2012a, p. 343).

Aquests models permeten ajustar qualsevol tipus de corba, encara que aquí només considerem el creixement lineal. Es tracta en realitat del cas més senzill de model de corbes de creixement.

Des del punt de vista del model estadístic, aquesta variabilitat individual de la ràtio d'increment dels ingressos amb el temps ve representada per un segon factor aleatori, que se suma al valor poblacional del coeficient de la variable que representa el temps (en el nostre cas, hem triat la variable "onada").

Suposem que β_1 és el quoeficient de la variable "onada", llavors la fórmula del model es representaria així:

$$y_{it} = (\beta_0 + \zeta_{0i}) + x_{1it}(\beta_1 + \zeta_{1i}) + x_{2it}\beta_2 + \ldots + x_{kit}\beta_k + \varepsilon_{it}$$

Els supòsits del model són els mateixos que en el cas del model d'intercept aleatori.

3.2.3. Descripció de les dades

Amb `(i)` cridem l'arxiu que conté les dades, `SELINGRE.DTA`.

```
use D:\SELINGRE.DTA,clear                                    (i)
```

Prèviament a qualsevol tractament longitudinal, deflactarem els ingressos d'acord amb l'IPC català

[15] En anglès s'utilitza sovint l'expressió models mixtos (*mixed models*) en lloc de models d'efectes aleatoris. L'expressió ve del fet que aquests models "barregen" factors fixos i factors aleatoris.

anual (segons l'INE, any base 2011) per tal d'anul·lar els efectes de la infació i guardarem els resultats com a `ingressos2 (ii)`.

```
gen ingressos2=ingressos * 1.260731981 if wave==2003                    (ii)
replace ingressos2=ingressos * 1.218412652 if wave==2004
replace ingressos2=ingressos * 1.172484142 if wave==2005
replace ingressos2=ingressos * 1.130991427 if wave==2006
replace ingressos2=ingressos * 1.098418278 if wave==2007
replace ingressos2=ingressos * 1.055264185 if wave==2008
replace ingressos2=ingressos * 1.053574251 if wave==2009
```

Seguidament, prendrem el logaritme dels ingressos per a utilitzar-lo com a variable dependent `(logingre) (iii)`. Aquesta transformació és convenient a causa de la forta assimetria positiva dels ingressos.

```
gen logingre=log(ingressos2)                                            (iii)
```

Amb la instrucció `(iv)` definim les dades com a dades de panel, en què el temps ve marcat pels anys transcorreguts des del 2003. La instrucció `xtdescribe (v)`, per la seva banda, ens permet examinar els patrons d'*attrition* de les dades (vegeu el quadre 12):

```
tsset id_ind onada                                                      (iv)
```

```
xtdescribe                                                              (v)
```

Quadre 12. Patrons d'*attrition* en les dades

```
  id_ind:  1.111e+11, 1.111e+11, ..., 1.211e+11              n =       1312
   onada:  0, 1, ..., 6                                      T =          7
           Delta(onada) = 1 unit
           Span(onada)  = 7 periods
           (id_ind*onada uniquely identifies each observation)

Distribution of T_i:   min      5%     25%       50%       75%     95%     max
                         1       1       3         5         7       7       7

     Freq.  Percent    Cum.  |  Pattern
 ---------------------------+---------
       611    46.57   46.57  |  1111111
       261    19.89   66.46  |  111....
       146    11.13   77.59  |  1......
       113     8.61   86.20  |  1111...
        94     7.16   93.37  |  11.....
        51     3.89   97.26  |  11111..
        36     2.74  100.00  |  111111.
 ---------------------------+---------
      1312   100.00          |  XXXXXXX
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Tenim doncs un total de 1.312 individus amb un màxim de set observacions cadascun, i un període d'espaiment (Delta) constant d'un any (el petit percentatge de patrons *returner* a la mostra original s'ha completat mitjançant imputació). En els patrons d'*attrition*, les ocasions en què l'individu no ha estat observat es representen mitjançant ".", mentre que aquelles en què sí que ho ha estat es

representen mitjançant "1". Una mica menys de la meitat dels individus (47%) ha completat les set rondes, però (per exemple) un 11% ja va caure a la primera ronda.

Ens trobem, per tant, davant de dades no equilibrades quant al nombre d'observacions. Tot i que aquesta és una característica més aviat indesitjable, la majoria de tècniques modernes d'anàlisi de dades longitudinals (i de forma particular, tal com estan implementades en Stata) permeten realitzar estimacions consistents a partir d'aquesta mena de dades –sota el supòsit que la pèrdua d'informació s'ha produït segons un patró MAR (Rabe-Hesketh i Skrondal, 2012a, p. 278).

Més concretament, en els models d'efectes aleatoris que fem servir aquí, tots els subjectes participen en les estimacions, amb independència del nombre d'onades amb què hi contribueix cadascun. Encara que les persones amb una sola observació no aporten cap informació sobre la variació intrasubjecte, sí que contribueixen a l'estimació dels efectes fixos (Singer i Willett, 2003, p. 148).

Un altre aspecte a considerar és el caràcter constant o variable de les variables del model. En les primeres, tota la variació de la variable té lloc entre subjectes (*between subjects*), mentre que en les segones la variació es reparteix entre subjectes i dintre els subjectes (*whithin subjects*). En un estudi longitudinal, naturalment, la variable dependent sempre canvia amb el temps.

Els comandaments `xtsum (vi)` i `xttab (vii a ix)` ens permeten examinar la variació intrasubjectes i entre subjectes per a variables quantitatives i qualitatives, respectivament.

```
xtsum logingre onada anys_treb tenure3                    (vi)
```

Quadre 13. Variació intraindividual i interindividual en covariables quantitatives

Variable		Mean	Std. Dev.	Min	Max	Observations
logingre	overall	7.375061	.5075494	4.751711	9.515231	N = 6317
	between		.4932269	5.212024	9.151002	n = 1312
	within		.2138551	5.79437	9.564045	T-bar = 4.81479
onada	overall	2.443565	1.964421	0	6	N = 6317
	between		1.136557	0	3	n = 1312
	within		1.751775	-.556435	5.443565	T-bar = 4.81479
anys_t~b	overall	23.43739	11.20169	1	54	N = 6317
	between		11.75121	2	54	n = 1312
	within		1.751775	20.43739	26.43739	T-bar = 4.81479
tenure3	overall	12.26057	10.09992	.0003785	49	N = 6317
	between		9.932237	.0833333	47	n = 1312
	within		2.835837	-14.16801	38.21536	T-bar = 4.81479

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

La variació dels ingressos entre subjectes (0.49) és més important que la variació intrasubjecte al llarg del període 2003-2009 (0.21), tal i com es pot veure al quadre 13. El mateix ocorre amb els anys treballats i la *tenure*. Això sembla lògic en una mostra que barreja individus molt heterogenis amb diverses característiques (moment del cicle vital, estudis, ocupació, etc). L'única variable en què la variació intraindividual és major que entre individus és l'onada. De fet, si les dades fossin equilibrades, la variació entre subjectes d'onada seria 0; però l'*attrition* fa que la mitjana d'onada no sigui la mateixa per a tots els individus.

```
xttab sexe                                                (vii)
```

Quadre 14. Variació intraindividual i interindividual en el factor "sexe"

	Overall		Between		Within
sexe	Freq.	Percent	Freq.	Percent	Percent
home	3625	57.38	749	57.09	100.00
dona	2692	42.62	563	42.91	100.00
Total	6317	100.00	1312	100.00	100.00
			(n = 1312)		

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

El quadre 14 ens mostra que la variable "sexe" no varia amb el temps. El 57% dels membres de la mostra han estat homes "en una o més observacions" i els individus a qu i això passa han romàs en la mateixa categoria (com a mitjana) un 100% de totes les seves observacions (columna "Within").

```
xttab estudis1                                                    (viii)
```

Quadre 15. Variació intraindividual i interindividual en el factor "nivell d'estudis"

	Overall		Between		Within
estudis1	Freq.	Percent	Freq.	Percent	Percent
'Primàri	902	14.28	229	17.45	98.06
'Primer	1038	16.43	222	16.92	98.57
'Segon c	750	11.87	160	12.20	98.07
'Segon c	781	12.36	172	13.11	94.72
'Educaci	664	10.51	140	10.67	96.81
'Estudis	2182	34.54	420	32.01	98.39
Total	6317	100.00	1343	102.36	97.69
			(n = 1312)		

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

En el cas dels estudis formals ("primària o menys", "primer cicle de secundària", "segon cicle de secundària tècnic", "segon cicle de secundària no tècnic", "postsecundària", "universitaris"), existeix una lleugera variació intrasubjecte, com ho indica el fet que la suma de la columna "Between" sigui superior al 100% (quadre 15). De tota manera, es tracta d'una característica essencialment estable, com ho indiquen els elevadissíms percentages de la columna "Within".

```
xttab intensitat2                                                 (ix)
```

Quadre 16. Variació intraindividual i interindividual en el factor "tipus de jornada"

	Overall		Between		Within
intensi~2	Freq.	Percent	Freq.	Percent	Percent
parcial	475	7.52	190	14.48	60.17
completa	5842	92.48	1241	94.59	96.51
Total	6317	100.00	1431	109.07	91.68
			(n = 1312)		

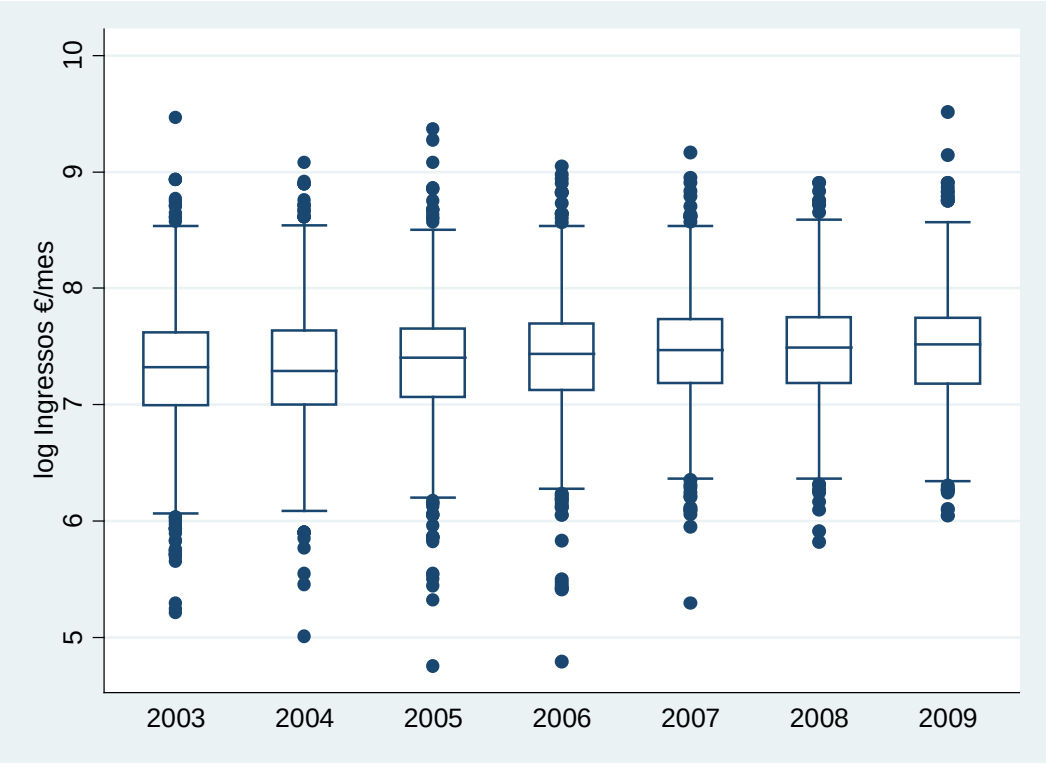
Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Finalment, el treball a temps complet o a temps parcial sí que presenta una major variabilitat intra-subjecte, propiciada fonamentalment per aquells que han treballat algun cop a temps parcial i que tendeixen a cambiar a treball a temps complet: de mitjana, un 60% dels seus registres indica que roman en situació de treball a temps parcial (quadre 16).

La figura 11 ens permet apreciar la distribució del logaritme dels ingressos (`log Ingressos`) i la seva evolució en el temps. Aquesta figura es crea a partir del comandament (`x`):

```
graph box logingre, over (wave) intensity (0) medtype(line) ytitle (x)
(log Ingressos €/mes)
```

Figura 11. Distribució i evolució dels ingressos en el període estudiat



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Tot i que es tracta dels ingressos deflactats, sembla que hi ha una lleugera tendència a l'alça dels mateixos, amb l'excepció de l'any 2004.

Ara bé, això no ens aclareix aspectes com ara l'*heterogeneïtat individual* i la possible evolució dels ingressos al nivel intraindividual. Podem fer-nos una idea d'això representant gràficament l'evolució en el temps d'una petita mostra d'individus. Presentarem els resultats desglossats per sexe (figura 12 i següents).

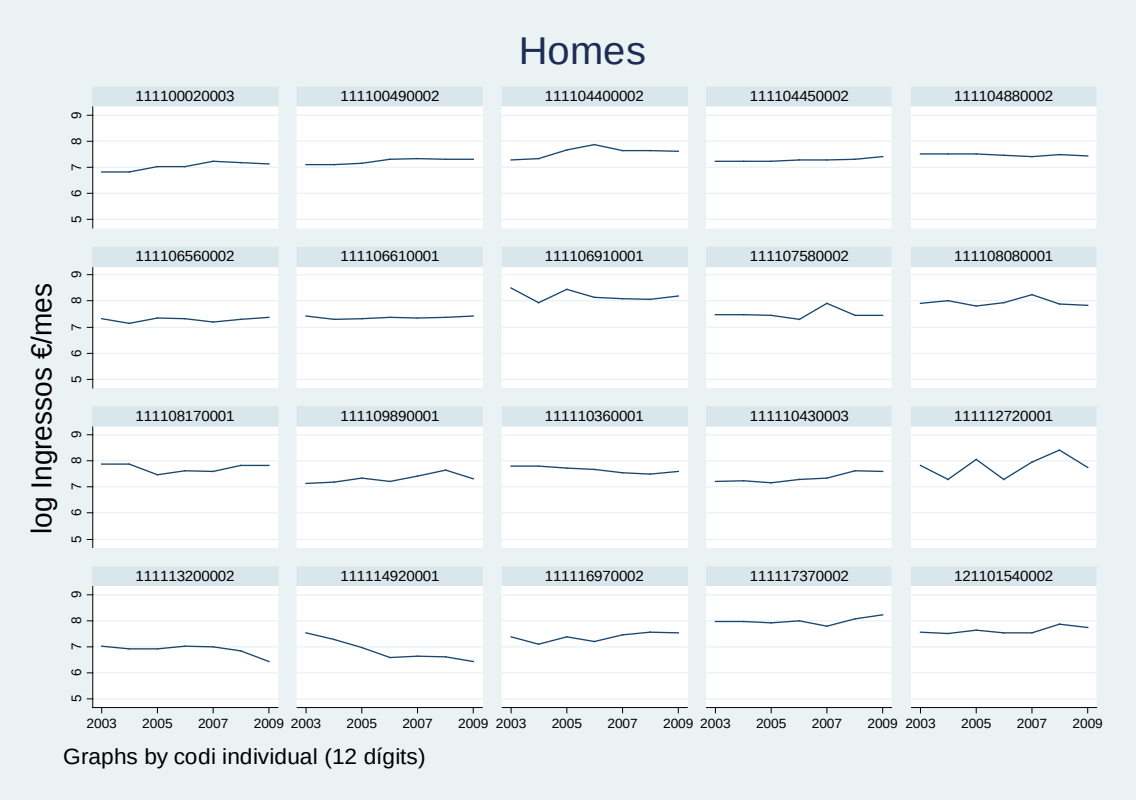
```
preserve                                                              (xi)


keep if max_onada==6 & sexe==0                                        (xii)
format logingre* %9.0g
sort id_ind onada
set seed 13210
generate r=uniform() if onada==0
egen num=rank (r) if r<.

egen number=mean(num), by (id_ind)                                    (xiii)
```

Atès que treballarem sobre una selecció de casos, el comandament `preserve (xi)` guarda l'arxiu de dades en la seva forma actual, que després recuperarem mitjançant el comandament `restore`. El conjunt dels comandaments `(xii)` i el comandament `(xiii)` permenten seleccionar una mostra de vint homes d'entre els que han completat les set rondes d'observacions, i que veiem representades individualment mitjançant el comandament següent:

```
twoway line logingre wave if number<=20, by(id_ind, title("Homes") )  (xiv)
ytitle(log Ingressos €/mes) xlab(2003 (2) 2009) xtitle("") ysca-
le(range(5 9))
```

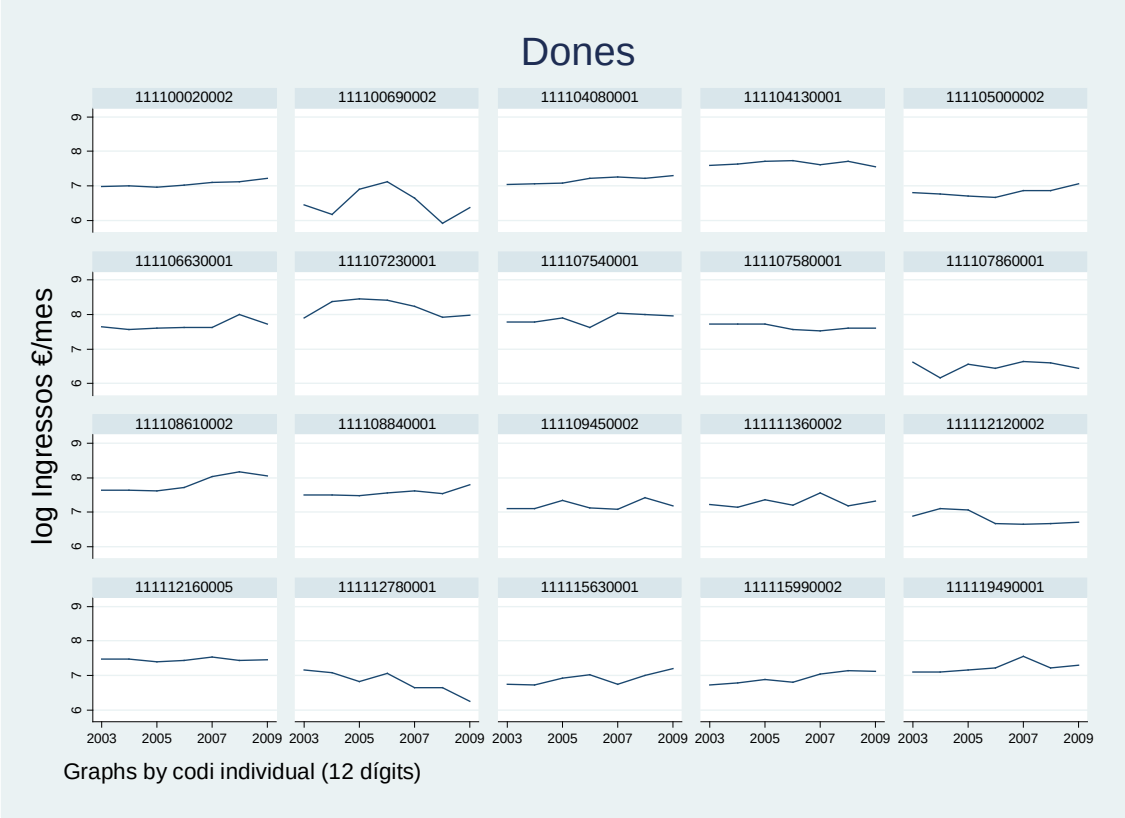
(Nota: obtenim els gràfics equivalents per a les dones simplement canviant el valor de sexe a 1 en el procedimient de selecció a partir de la instrucció (xii), i modificant altres aspectes pertinents del format, com ara el títol –aquests comandaments s'ometen en el text.)

Figura 12. Evolució dels ingressos per una mostra de vint homes *completers*



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Figura 13. Evolució dels ingressos per una mostra de dones *completers*



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Com es veu, existeix força heterogeneïtat individual en el nivell inicial d'ingressos, tal pel que fa als homes com a les dones. Aquesta heterogeneïtat és la que afronta el model A mitjançant la introducció d'*intercept* aleatoris.

Tanmateix, el model A continua suposant que els coeficients de les variables predictores, un cop s'ha condicionat per les característiques individuals, són els mateixos per a tots els individus i les diferències es deuen a l'atzar.

Les figures 12 i 13, en canvi, mostren clarament que les diferències no se situen tan sols al nivell del nivell d'ingressos inicials, sinó que també hi ha diferències en la manera com canvien al llarg del temps. Encara que en molts individus s'observa una lleugera tendència a l'alça, alguns presenten una trajectòria irregular o fins i tot descendent.

Aquesta segona font d'heterogeneïtat és la que adreça el model B1. Aquest no tan sols permet diferències entre els individus en el nivell inicial d'ingressos, sinó que admet també diferències en l'evolució dels ingressos en el període.

Naturalment, la idea d'heterogeneïtat en el creixement dels ingressos només té sentit si existeix realment una evolució significativa dels ingressos al llarg del període, més enllà de la variació aleatòria. Els models ens confirmaran aquest particular, però en termes descriptius podem fer una primera aproximació de la tendència de conjunt dibuixant l'anomenat *spaghetti plot* (comandament xv) amb una funció lineal ajustada a les dades de tres-cents individus *completers,* que permet apreciar millor la tendència de conjunt:

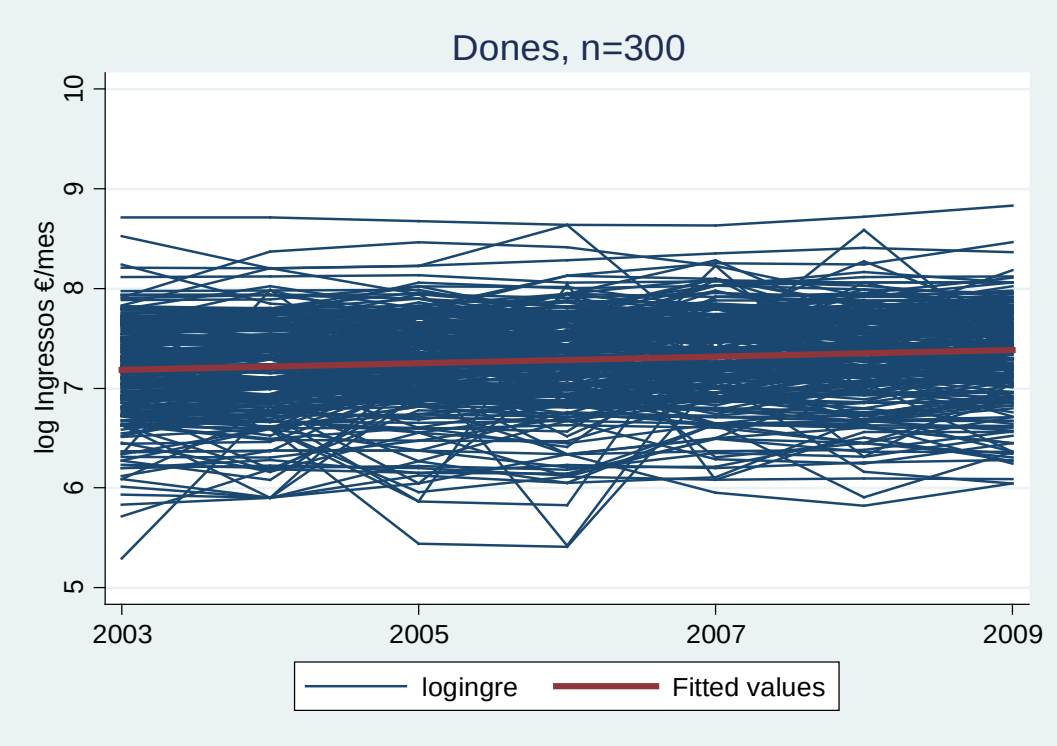
```
twoway (line logingre wave,connect(L)) (lfit logingre wave,connect(L)  (xv)
lwidth(thick))  if  number<=300,  ytitle(log  Ingressos  €/mes)  tit-
le("Homes, n=300") xlab(2003 (2) 2009) xtitle("")
```

S'aprecia una lleugera tendència ascendent dels ingressos (figura 14), una mica més acusada en el cas de les dones (figura 15). La mitjana dels ingressos és tanmateix sempre major en el cas dels homes.

Figura 14. Evolució dels ingressos per una mostra de tres-cents homes *completers* i ingressos mitjans ajustats

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Figura 15. Evolució dels ingressos per una mostra de tres-centes dones *completers* i ingressos mitjans ajustats



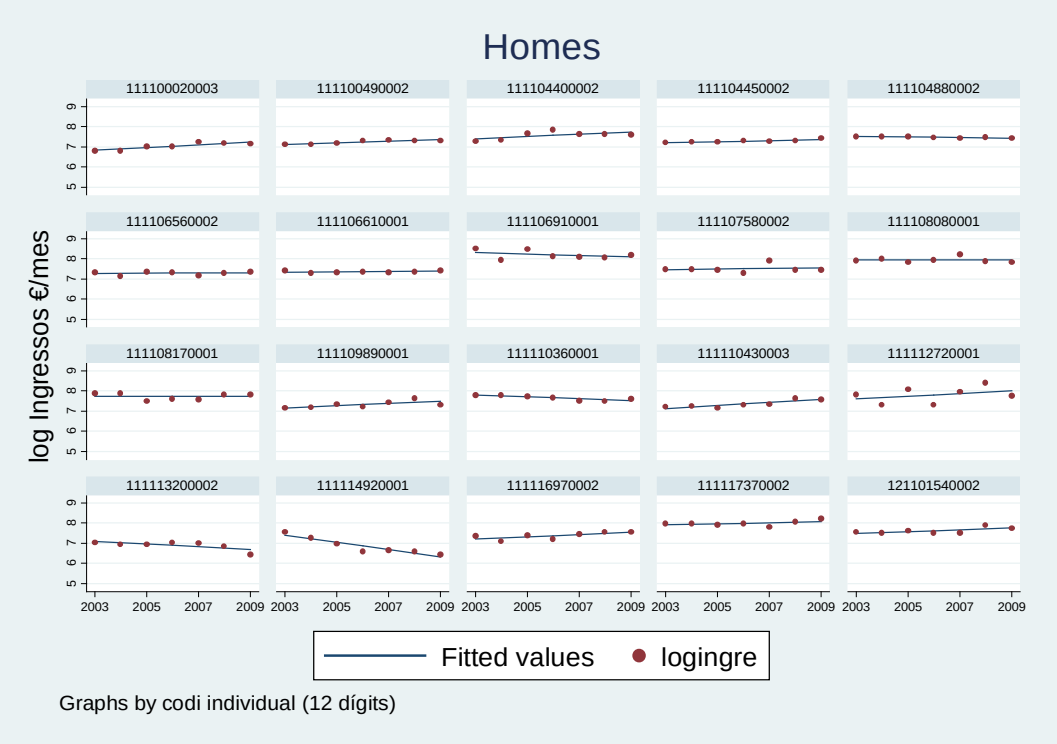
Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Finalment presentem l'evolució dels salaris i la seva trajectòria lineal ajustada d'una mostra de vint homes *completers* del panel (comandament `xvi`; figura 16). Amb alguna excepció, l'evolució dels seus ingressos sembla que es pugui ajustar raonablement bé mitjançant una recta:

```
graph twoway (lfit logingre wave)(scatter logingre wave) if num- (xvi)
ber<=20,    by(id_ind,  title("Homes"))  ytitle(log  Ingressos  €/mes)
xlab(2003 (2) 2009) xtitle("") yscale(range(5 9))
```

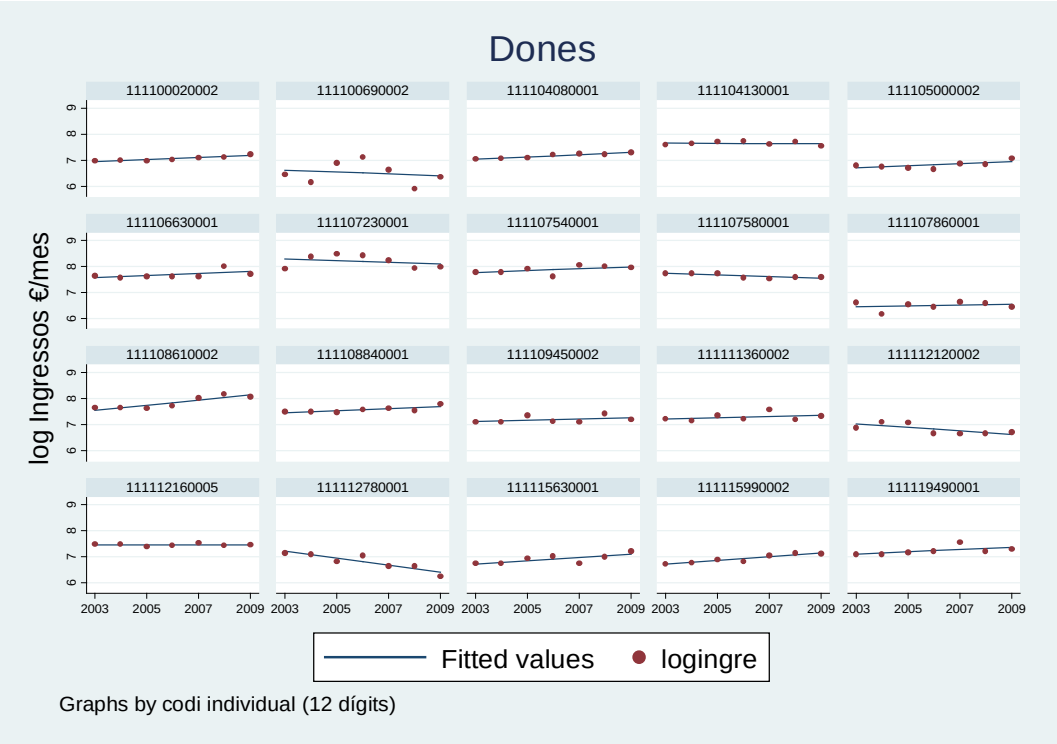
Arribem a conclusions molt similars en el cas de les dones (vegeu la figura 17).

Figura 16. Evolució dels salaris i la trajectòria lineal ajustada en funció de l'onada



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Figura 17. Evolució dels salaris i la trajectòria lineal ajustada en funció de l'onada



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

3.2.4. Execució dels models

3.2.4.1. Execució dels models referits a la qüestió A.

Model A1. Regressió ordinària amb errors estàndard robustos

Les variables seleccionades per al model tenen realment un efecte sobre els ingressos? Una primera manera d'examinar-ho és fer servir un model de regressió amb estimació pel mètode de mínims quadrats (xviii).

De forma prèvia convertim la variable politòmica "nivell d'estudis" en un seguit de variables dummy (comandament xvii), ja que l'Stata només admet en la modelització variables contínues o dicotòmiques, i seguidament donem a les variables creades un nom més descriptiu (segona i tercera línies del comandament xvii).

```
tab estudis1,gen(estud)                                                    (xvii)
rename (estud2 estud3 estud4 estud5 estud6) (Ci1Sec  Ci2SecT  Ci2SecNT
Post2NU Univ)

regress logingre anys_treb onada tenure3 sexe Ci1Sec    Ci2SecT  (xviii)
Ci2SecNT Post2NU Univ intensitat2, vce(cluster id_ind)
```

Quadre 17. Estimacions del model A1

```
Linear regression                               Number of obs =    6317
                                                F( 10,  1311) =  109.29
                                                Prob > F      =  0.0000
                                                R-squared     =  0.3833
                                                Root MSE      =  .39889

                          (Std. Err. adjusted for 1312 clusters in id_ind)
------------------------------------------------------------------------------
             |               Robust
    logingre |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
   anys_treb |   .0097578   .0013028     7.49   0.000     .0072019    .0123136
       onada |   .0162335   .0026443     6.14   0.000      .011046     .021421
     tenure3 |   .0030116   .0012838     2.35   0.019     .0004931    .0055302
        sexe |  -.2956704   .0213774   -13.83   0.000     -.337608   -.2537328
      Ci1Sec |   .2244826   .0425832     5.27   0.000     .1409438    .3080213
      Ci2SecT |   .2258709   .0433995     5.20   0.000     .1407307     .311011
    Ci2SecNT |   .4221617   .0434699     9.71   0.000     .3368835    .5074398
     Post2NU |    .389035   .0432581     8.99   0.000     .3041724    .4738976
        Univ |   .6955229     .03825    18.18   0.000      .620485    .7705607
  intensitat2 |   .4984804   .0370354    13.46   0.000     .4258252    .5711355
       _cons |    6.33774   .0569455   111.29   0.000     6.226025    6.449454
------------------------------------------------------------------------------
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

El model en el seu conjunt és clarament significatiu en relació amb el model basal i explica un 38% del total de la variança dels ingressos. El valor de la constant ens indica que un home amb estudis primaris o menys (categoria de referència), sense cap experiència laboral prèvia i treballant a temps parcial tindria el 2003 uns ingressos mitjans de $e^{6.33774} = 565.5$ euros (IC 534.2-598.7). Certament

aquest és un valor poc creïble en referència a la població general, però com ja hem advertit, en aplicar determinats criteris de selecció hem creat un biaix de selecció considerable.

Els efectes de les variables explicatives van en el sentit previsible. El fet de ser dona, per exemple, augmenta *ceteris paribus* els ingressos esperats en $e^{-0.2956704}-1=-0.256$, és a dir, els redueix un 25,6%. Tenir un nivell d'estudis universitaris els augmenta en $e^{0.6955229}-1=1.005$, és dir pràcticament els dobla en relació amb tenir estudis primaris o menys. Cada any treballat augmenta $e^{0.0097578}-1=0.0098$, és a dir aproximadament un 1% els ingresos esperats.

Cal destacar que tant els anys treballats com l'onada tenen un efecte significatiu (i de magnitud similar), la qual cosa suggereix que al llarg del període 2003-2009 es va produir un creixement anual dels ingressos entorn de l'1,6%.

A partir del model que hem ajustat, podem calcular els residus i l'autocorrelació existent entre ells (`xix`).

```
predict res, residuals                                          (xix)
```

El comandament `reshape (xx)` transforma l'arxiu de format *long* a *wide* per permetre calcular la correlació entre els residus de cada onada (quadre 18).

```
preserve                                                        (xx)
keep id_ind res onada
reshape wide res, i(id_ind) j(onada)
```

Quadre 18. Transformació de les dades de format *long* a format *wide*

```
(note: j = 0 1 2 3 4 5 6)

Data                              long   ->   wide
-----------------------------------------------------------------------------
Number of obs.                    6317   ->    1312
Number of variables                  3   ->       8
j variable (7 values)            onada   ->   (dropped)
xij variables:
                                   res   ->   res0 res1 ... res6
-----------------------------------------------------------------------------
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

La variança dels residus és bastant similar per cada onada, és a dir es compleix aproximadament la condició d'homoscedasticitat (quadre 19), comandament `(xxi)`

```
tabstat res*, statistics(sd) format (%3.2f)                     (xxi)
```

Quadre 19. Desviació típica de residus. Model A1

stats	res0	res1	res2	res3	res4	res5	res6
sd	0.39	0.39	0.42	0.42	0.39	0.38	0.39

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

```
pwcorr res*                                                    (xxii)
restore
```

Quadre 20. Correlacions entre els residus. Model A1

	res0	res1	res2	res3	res4	res5	res6
res0	1.0000						
res1	0.7250	1.0000					
res2	0.6722	0.6709	1.0000				
res3	0.5996	0.6306	0.6960	1.0000			
res4	0.6012	0.5954	0.6546	0.6415	1.0000		
res5	0.6188	0.5837	0.6578	0.6087	0.7064	1.0000	
res6	0.6150	0.5851	0.6744	0.6241	0.6919	0.7533	1.0000

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

En aquesta matriu (quadre 20, comandament `xxii`), res0 representa els residus de l'onada 0, res1 els de l'onada 1, etc. Clarament no es compleix la condició d'independència dels residus. Veiem, en canvi, que existeix una forta correlació entre aquests, que tendeix a disminuir lleugerament a mesura que comparem anys més allunyats. Recordem que la no-existència de correlació entre els residus és un dels supòsits bàsics de la tècnica de la regressió, i que per tant és un problema que cal tractar.

Les estimacions obtingudes mitjançant aquest primer model A1 són consistents (és a dir convergeixen cap als valors veritables quan la mostra és gran), mentre que l'opció `vce(cluster id_ind)` permet calcular també els errors estàndard dels estimadors de forma consistent, tenint en compte l'esmentada correlació entre els residus d'un mateix individu.

La principal limitació d'aquesta aproximació per mínims quadrats generalitzats és que assumeix que els valors perduts (i en particular, els originats per l'*attrition*) en la variable dependent en el moment t, són independents del valor de la resposta en el moment t-1.

En canvi, l'estimació pel mètode de màxima versemblança que empren les aproximacions més modernes, i en particular els models d'efectes aleatoris, no necessita realitzar aquesta assumpció. Per aquesta raó, presenten moltes més garanties quan (com és habitual) les dades son de caràcter no equilibrat (Rabe-Hesketh i Skrondal, 2012a, p. 242, 279-282).

*Model A2. Model d'*intercept *aleatori*

El comandament `xtmixed (xxiii)` permet ajustar models lineals d'efectes aleatoris (quadre 21). La part aleatòria del model comença després de la doble barra, i la instrucció `mle` demana que es faci l'estimació pel mètode de màxima versemblança.

```
xtmixed logingre anys_treb onada tenure3 sexe Ci1Sec    Ci2SecT (xxiii)
Ci2SecNT Post2NU Univ intensitat2 || id_ind:, mle
```

Quadre 21. Estimacions del model A2

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -1227.1987
Iteration 1:   log likelihood = -1227.1987

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      6317
Group variable: id_ind                          Number of groups   =      1312

                                                Obs per group: min =         1
                                                               avg =       4.8
                                                               max =         7


                                                Wald chi2(10)      =   1462.61
Log likelihood = -1227.1987                     Prob > chi2        =    0.0000
```

logingre	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
anys_treb	.0093419	.0010388	8.99	0.000	.007306	.0113779
onada	.0130817	.0018966	6.90	0.000	.0093645	.0167989
tenure3	.0031188	.000814	3.83	0.000	.0015234	.0047141
sexe	-.3362966	.0207543	-16.20	0.000	-.3769743	-.2956189
Ci1Sec	.202331	.0350054	5.78	0.000	.1337217	.2709404
Ci2SecT	.2530845	.0377475	6.70	0.000	.1791008	.3270683
Ci2SecNT	.4050108	.0366701	11.04	0.000	.3331387	.476883
Post2NU	.4169482	.0405763	10.28	0.000	.3374201	.4964764
Univ	.7131388	.0328889	21.68	0.000	.6486776	.7775999
intensitat2	.2995152	.0188491	15.89	0.000	.2625716	.3364587
_cons	6.535563	.0431697	151.39	0.000	6.450952	6.620174

Random-effects Parameters	Estimate	Std. Err.	[95% Conf.	Interval]
id_ind: Identity				
sd(_cons)	.3374292	.0076301	.3228011	.3527202
sd(Residual)	.2316499	.0023253	.2271369	.2362525

```
LR test vs. linear regression: chibar2(01) =  3849.98 Prob >= chibar2 = 0.0000
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Els estimadors de la part fixa del model són similars als del model A1, amb l'excepció del treball a temps complet, que ha disminuït notablement el seu efecte.

A la part aleatòria del model, `sd(_cons)` és la desviació típica de la part aleatòria de l'*intercept* ζ_{0i} (desviacions respecte a l'*intercept* poblacional), val 0.337 i representa la variabilitat entre individus no explicada pel model. La desviació típica dels residus ε_{it} és menor, val `sd(_cons)`= 0.232 i representa la variabilitat intraindividual no explicada pel model. Ambdues magnituds són significativament diferents de zero, la qual cosa indica que podem continuar incrementant la capacitat explicativa del model a partir de la introducció de nous factors predictors.

A partir d'aquestes desviacions típiques podem calcular el coeficient d'autocorrelació:

$$\rho = \frac{\hat{\psi}}{\hat{\psi} + \hat{\theta}} = \frac{0.337^2}{0.337^2 + 0.232^2} = 0.680$$

Això significa que el 68,0% de la variança dels ingressos no explicada per les covariables es deu a característiques no observades dels subjectes, invariables en el temps (Rabe-Hesketh i Skrondal, 2012a, p. 250).

3.2.4.2. Execució dels models referits a la qüestió B.

Model B1 de coeficients aleatoris / corba de creixement

En el model anterior hem vist que el número d'onada té un efecte significatiu i positiu sobre els ingressos, independent de l'increment de l'experiència (anys treballats) i que interpretem com un efecte del període sobre els ingressos.

Ara bé, l'efecte del període és realment igual per a tots els individus de la mostra? En el model A2 d'intercept aleatori permetíem valors inicials (intercepts) diferents específics de cada individu, però suposàvem que els coeficients i per tant el pendent de la recta (en realitat, hiperplà, ja que ajustàvem diverses covariables) era el mateix per a tots els individus. Aquest supòsit és irrealista, ja que en els descriptius (figures 16 i 17) hem pogut veure casos d'individus amb pendent negatiu.

Ara permetrem que no tan sols l'intercept del model variï aleatòriament, sinó també el coeficient de la variable "onada". Això significa que permetem també certa variabilitat individual quant a la ràtio d'increment dels ingressos al llarg del període d'observació (ara bé, continuem suposant que la forma funcional de creixement dels ingressos és lineal).

La introducció d'un coeficient aleatori per a la variable "onada" permetrà explicar millor la variabilitat dels ingressos al nivell intraindividual, tot reduint la variabilitat residual existent a aquest nivell.

Ajustem aquest nou model (quadre 22) amb Stata mitjançant la instrucció `xxiv`:

```
xtmixed logingre onada anys_treb tenure3 sexe Ci1Sec Ci2SecT (xxiv)
Ci2SecNT Post2NU Univ intensitat2 || id_ind: onada, cov(un) mle
```

Veiem que l'única novetat és la introducció de la variable "onada" en la part aleatòria del model, i que permetem que la covariança entre els dos coeficients aleatoris (*intercept* i pendent) s'estimi lliurement cov(un).

Quadre 22. Estimacions del model B1

```
Performing EM optimization:

Performing gradient-based optimization:

Iteration 0:   log likelihood = -1176.8462
Iteration 1:   log likelihood = -1176.2355
Iteration 2:   log likelihood = -1176.2354
Iteration 3:   log likelihood = -1176.2354

Computing standard errors:

Mixed-effects ML regression                     Number of obs      =      6317
Group variable: id_ind                          Number of groups   =      1312

                                                Obs per group: min =         1
                                                               avg =       4.8
                                                               max =         7


                                                Wald chi2(10)      =   1325.62
Log likelihood = -1176.2354                     Prob > chi2        =    0.0000
```

logingre	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada	.0131957	.002257	5.85	0.000	.0087721	.0176194
anys_treb	.009511	.0010487	9.07	0.000	.0074557	.0115663
tenure3	.0029484	.0008484	3.48	0.001	.0012856	.0046112
sexe	-.3401856	.0208123	-16.35	0.000	-.3809769	-.2993944
Ci1Sec	.2016411	.0350322	5.76	0.000	.1329793	.2703029
Ci2SecT	.2507165	.0377866	6.64	0.000	.1766562	.3247768
Ci2SecNT	.4119326	.0368929	11.17	0.000	.3396238	.4842414
Post2NU	.4119185	.0407702	10.10	0.000	.3320105	.4918266
Univ	.71293	.0330251	21.59	0.000	.6482021	.7776579
intensitat2	.2842782	.0190367	14.93	0.000	.2469669	.3215894
_cons	6.549194	.0434017	150.90	0.000	6.464128	6.63426

```
------------------------------------------------------------------------------
  Random-effects Parameters |   Estimate   Std. Err.      [95% Conf. Interval]
----------------------------+-------------------------------------------------
id_ind: Unstructured        |
                  sd(onada) |    .0380274   .0025464       .0333503    .0433605
                  sd(_cons) |    .3494833   .0085447       .3331309    .3666384
           corr(onada,_cons) |   -.2508728   .0564684       -.357881   -.1373606
----------------------------+-------------------------------------------------
                sd(Residual) |    .2194248   .0024448       .2146851    .2242692
------------------------------------------------------------------------------
LR test vs. linear regression:       chi2(3) =  3951.91   Prob > chi2 = 0.0000

Note: LR test is conservative and provided only for reference.

.
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Pràcticament no existeixen diferències en les estimacions de la part fixa del model entre el model A2 i el B1. En la part aleatòria del model, a part de la desviació típica de l'*intercept* i la residual, apareix ara la desviació típica del pendent `sd(onada)`= .0380274 i la correlació entre els dos coeficients aleatoris *intercept* i pendent `corr(onada,_cons)`= -.2508728.

Es comprova que la variància deguda a diferències en l'estatus inicial (*intercept*) és molt major que la variància deguda a la raó de canvi en l'onada. Continua havent-hi variança explicable al nivell intraindividual (la variança residual és diferent de zero). Per explicar-la, hi podríem afegir nous factors aleatoris corresponents a altres variables.

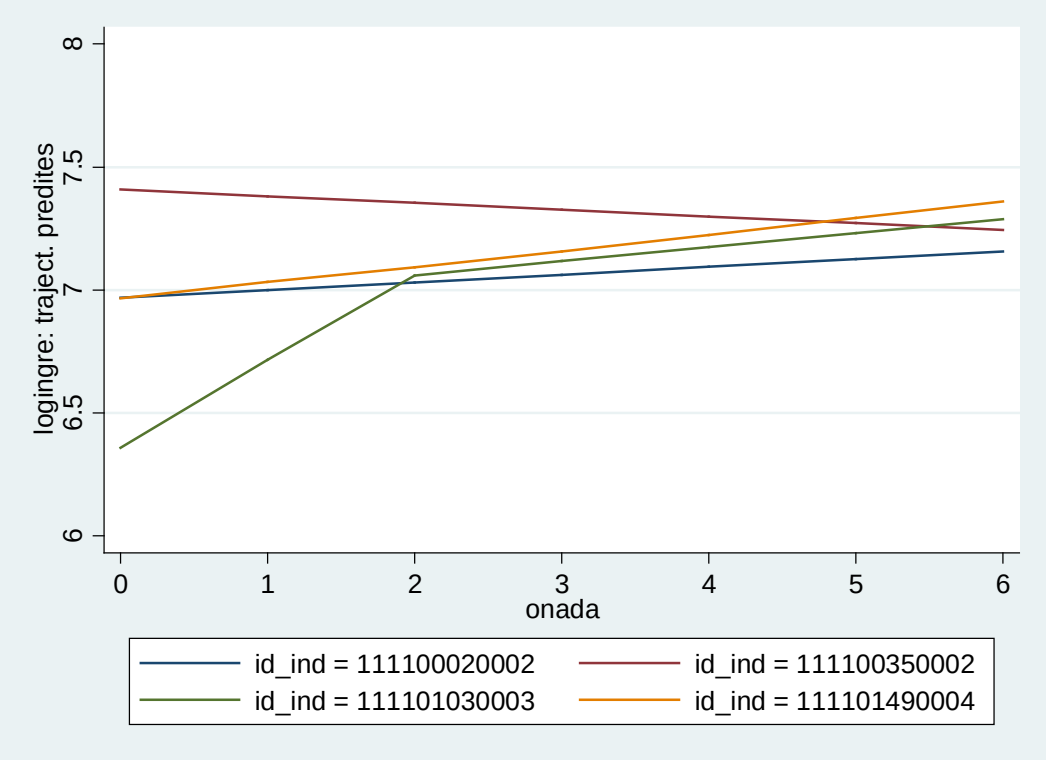
Atès que l'efecte poblacional estimat de la variable "onada" és 0.0131957, l'efecte individual oscil□larà (amb un 95% de confiança) entre 0.0131957± 1.96 x 0.0380274, és a dir, entre -0.061338 i 0.0877294 "logeuros" anuals. Per tant, veiem que el pendent dels ingressos serà negatiu (és a dir, disminuiran) per a moltes persones en el conjunt del període.

El comandament `xxv` crea un gràfic amb l'evolució predita dels ingressos per a quatre individus seleccionats:

```
predict traj, fitted                                                    (xxv)
xtline  traj  if  id_ind=111101490004  ||  id_ind=111105850001  ||
id_ind=111115110002 || id_ind=111100690002, overlay t(onada) i(id_ind)
scheme(s2mono)
```

La figura 18 mostra la trajectòria individual predita pel model per a quatre individus. Observem que gràcies a l'efecte aleatori introduït per a onada, el pendent és diferent per a cada individu, encara que la forma funcional de l'evolució dels ingressos és sempre una recta.

Figura 18. Trajectòries predites pel model B1 de quatre individus seleccionats



Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

La correlació entre *intercept* i *slope* és moderada i negativa, -0.25, la qual cosa ens indica (Singer i Willett, 2003, p. 100) que l'increment del salari i el seu valor inicial estan negativament associats al nivell intraindividual, és a dir: creixen més ràpidament els salaris més baixos. Això té cert sentit si resulta raonable suposar que existeix un cert "efecte sostre" en els ingressos: els de les persones que es troben més prop d'aquest sostre creixen menys ràpidament que els de les persones que han començat l'observació lluny d'aquest.

Taula 4. Components de la variança* i ajust en els models d'*intercept* aleatori (A2) i *intercept* + pendent aleatori (B1)

		A2	B1
Nivell 1 - Intraindividual		0.2316	0.2194
Nivell 2 - Entre individus			
	Intercept	0.3374	0.3495
	Pendent	---	0.0380
	Correlació I/P	---	-0.2509
Ajust			
	Log- likeli-hood	-1227.1987	-1176.2354

* Els components s'expressen en desviacions típiques.
Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

La inclusió del pendent aleatori per a "onada" ha permès disminuir la variança intraindividual no explicada en $\frac{0.2316^2 - 0.2194^2}{0.2316^2} = 0.103$, és a dir un 10,3%. En canvi, la variança de l'*intercept* ha

augmentat aproximadament un 7%. Aquesta situació sorprenent es dóna quan gairebé tota la variança es concentra o bé entre individus o bé dintre individus (Singer i Willett, 2003, p. 104).

3.2.5. Nota sobre els models d'efectes fixos

No podem cloure aquest cas pràctic sense una referència als models d'efectes fixos, que són utilitzats amb preferència en l'àmbit de la ciència econòmica per modelitzar situacions similars a la que ens hem plantejat aquí.

Els models d'efectes fixos, igual que els d'efectes aleatoris, estimen els coeficients dels factors de manera condicionada al subjecte, però mentre en els segons l'efecte del subjecte s'introdueix com una variable aleatòria, en els models d'efectes fixos s'introdueix com una variable més en la part fixa del model.

Un dels supòsits fonamentals dels models d'efectes aleatoris és que la part aleatòria del model no està correlacionada amb les variables regressores, tant si aquestes són observades com no observades (aquest supòsit està implícit en les assumpcions 1 i 2 del model A2).

Si el supòsit esmentat no es compleix, es diu que les variables regressores són endògenes (Baum, 2006). Si existeix endogeneïtat, les estimacions dels models d'efectes aleatoris seran esbiaixades, mentre que aquest problema no es presenta en els models d'efectes fixos. Com a contrapartida, les estimacions dels models d'efectes aleatoris són més precises en absència d'endogeneïtat.

A la pràctica, és molt habitual que les variables en l'àmbit de les ciències socials presentin problemes d'endogeneïtat. Per aquesta raó, concretament en econometria s'aplica sistemàticament el test de Hausman, que permet decidir sobre la presència d'endogeneïtat. Existeixen diverses estratègies per redreçar aquest problema, però en econometria generalment es prefereix optar per models d'efectes fixos.

Atès que controlen perfectament totes les fonts de variació interindividual, els models d'efectes fixos es concentren únicament en la variació intraindividual. Això mateix fa que no puguin estimar directament els coeficients de les variables constants en el temps, cosa que probablement explica la seva infrautilització en sociologia.

Anem a comparar els resultats d'un model d'*intercept* aleatori amb un model d'efectes fixos. Només considerarem les variables onada, *tenure* i intensitat, ja que la resta són constants al nivell individual –el nivell d'estudis presenta molt poca variabilitat i provoca estimacions molt imprecises en el model d'efectes fixos (model A3). Vegeu el quadre 23, comandament `xvi`:

```
xtreg logingre onada tenure3 intensitat2, fe                    (xxvi)
```

Quadre 23. Estimacions del model A3 (efectes fixos)

```
Fixed-effects (within) regression               Number of obs      =      6317
Group variable: id_ind                          Number of groups   =      1312

R-sq:  within  = 0.0677                         Obs per group: min =         1
       between = 0.2292                                        avg =       4.8
       overall = 0.1469                                        max =         7

                                                F(3,5002)          =    121.17
corr(u_i, Xb)  = 0.2481                         Prob > F           =    0.0000

------------------------------------------------------------------------------
    logingre |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      onada |   .0230703   .0017703    13.03   0.000     .0195998    .0265407
    tenure3 |    .001965   .0010931     1.80   0.072    -.0001779    .0041079
intensitat2 |    .248652   .0209716    11.86   0.000     .2075385    .2897654
      _cons |   7.064641   .0232841   303.41   0.000     7.018994    7.110288
------------------------------------------------------------------------------
    sigma_u |  .46172994
    sigma_e |  .23202545
        rho |  .79839068   (fraction of variance due to u_i)
------------------------------------------------------------------------------
F test that all u_i=0:     F(1311, 5002) =    15.92          Prob > F = 0.0000
```

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

```
estimates store fi                                                    (xxvii)
```

La instrucció (xxvii) guarda les estimacions del model d'efectes fixos per realitzar després el test de Hausman (xxviii). El mateix model amb efectes aleatoris (quadre 24):

```
xtreg logingre onada tenure3 intensitat2, re                          (xviii)
```

Quadre 24. Estimacions del model A3 (efectes aleatoris)

```
Random-effects GLS regression                   Number of obs      =      6317
Group variable: id_ind                          Number of groups   =      1312

R-sq:  within  = 0.0658                         Obs per group: min =         1
       between = 0.2240                                        avg =       4.8
       overall = 0.1538                                        max =         7

                                                Wald chi2(3)       =    595.97
corr(u_i, X)   = 0 (assumed)                    Prob > chi2        =    0.0000
```

logingre	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada	.0224441	.0017217	13.04	0.000	.0190697	.0258185
tenure3	.0054079	.0008181	6.61	0.000	.0038044	.0070113
intensitat2	.3325005	.0194318	17.11	0.000	.2944149	.3705862
_cons	6.918392	.0231415	298.96	0.000	6.873036	6.963749
sigma_u	.41303737					
sigma_e	.23202545					
rho	.76012812	(fraction of variance due to u_i)				

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

```
estimates store ri2                                                    (ixxx)
```

I finalment realitzem el test de Hausman (quadre 25):

```
hausman fi ri2                                                         (xxx)
```

Quadre 25. Test de Hausman

	Coefficients (b) fi	(B) ri2	(b-B) Difference	sqrt(diag(V_b-V_B)) S.E.
onada	.0230703	.0224441	.0006262	.0004119
tenure3	.001965	.0054079	-.0034429	.0007249
intensitat2	.248652	.3325005	-.0838486	.0078874

b = consistent under Ho and Ha; obtained from xtreg
B = inconsistent under Ha, efficient under Ho; obtained from xtreg

Test: Ho: difference in coefficients not systematic

chi2(3) = (b-B)'[(V_b-V_B)^(-1)](b-B)
= 237.32
Prob>chi2 = 0.0000

Font: Panel de Desigualtats Socials a Catalunya, 2003-2009.

Com és habitual, el test de Hausman surt significatiu, la qual cosa suggereix que una o més de les covariables examinades són endògenes.

3.3. Model de risc a temps discret: estudi sobre la pèrdua de llars en la mostra del PaD

Aquest apartat explica l'ajust d'un model de la família de les anàlisis de supervivència (model de risc a temps discret), aplicat a l'estudi dels factors que determinen l'*attrition* o abandonament de la participació en el PaD per les famílies de la mostra inicial.

El programari utilitzat ha estat Stata/SE® v.12.

3.3.1. Plantejament de l'estudi

Què és el desgast mostral?

Per desgast mostral (attrition) s'entén generalment la no-resposta d'unitat quan aquesta ocorre en el context d'un disseny longitudinal, és a dir, la impossibilitat d'entrevistar o mesurar una unitat mostral prèviament seleccionada, en qualsevol de les observacions programades al llarg del període d'observació (Peracchi, 2002).

En la definició convencional d'attrition s'inclouen tant els casos que rebutjen respondre l'enquesta en la ronda corresponent com aquells amb qui no és possible contactar.

En el cas de dissenys d'enquestes de panel a llars com és el PaD, la font de l'attrition pot ser tant la pèrdua d'individus com la de la totalitat de la llar.

Per què és important analitzar el desgast mostral?

El desgast mostral suposa sempre una pèrdua de potència de l'estudi i pot induir un biaix en la mostra, si la pèrdua d'unitats d'anàlisi no es produeixen a l'atzar sinó que afecta de forma selectiva a subjectes amb característiques específiques vinculades als objectius de l'estudi. En aquest cas, els resultats de l'estudi i les estimacions corresponents seran esbiaixats. D'altra banda, la reducció de la mida mostral també pot dificultar l'aplicació de tècniques estadístiques per analitzar subgrups específics de població (Lemay, 2009). Tots aquests aspectes fan especialment rellevant l'estudi de les causes del desgast mostral en els estudis de panel.

Quins són els determinants del desgast mostral?

El desgast mostral es pot explicar per factors molt diversos, com ara el nivell socioeconòmic de la llar o la seva ubicació geogràfica, el nivell d'estudis de l'informant principal o la seva situació laboral. La significació i el pes dels factors presenta gran variabilitat en funció de les característiques del panel o de la regió o el país en estudi.

Objectiu de l'estudi

En aquest exemple ens proposem estudiar l'evolució de l'attrition de les llars en el PaD, així com determinar els factors que hi poden estar relacionats mitjançant un model de risc a temps discret (discrete time hazard model). Hem seleccionat com a variables potencialment predictores l'idioma de preferència a l'hora de realitzar l'entrevista i la ubicació del municipi de residència de la llar en l'eix rural-urbà. Aquesta darrera variable ha estat socialitzada al PaD per la Fundació del Món Rural – Ignasi Aldomà i consta de quatre categories, establertes en funció de la població total del municipi i la seva densitat (Fundació del Món Rural i Aldomà Buixadé, 2009).

Operacionalització de l'attrition i selecció de casos

En aquest estudi farem servir una definició més restrictiva que la donada inicialment del desgast mostral, per la qual entendrem la impossibilitat d'entrevistar una unitat mostral prèviament entrevistada en una o més rondes incloent necessàriament la primera, en qualsevol de les observacions programades al llarg del període d'observació. Excloem per tant la no-resposta inicial (Taris, 2000,

p. 20).

D'acord amb aquesta definició, l'anàlisi del desgast mostral s'ha limitat a les llars efectivament entrevistades en la primera ronda del PaD (2002). Se n'han exclòs les llars incorporades en operacions posteriors de remostreig a partir del 2007, així com les llars de nova creació. Igualment s'han exclòs quatre llars amb valor perdut en la variable "idioma").

El desgast mostral s'ha definit com un esdeveniment amb dos estats possibles (permanència en el panel i abandonament de la llar)[16] i no repetible (és a dir, només s'estudia l'ocurrència de la primera no-observació). No s'han considerat, per tant, les reentrades de llars en el panel després d'un primer abandonament (escasses, d'altra banda).

La mostra així seleccionada inclou 1.987 llars i 9.416 registres acumulats al llarg de les vuit onades analitzades (del 2002 al 2009).

3.3.2. Caracterització del model de riscos a temps discret

L'interès de la recerca que hem plantejat se centra evidentment en l'ocurrència d'un primer abandonament del panel. Però també resulta de gran interès el moment en què això succeeix. Des del punt de vista de la gestió del panel, certament no és el mateix que una llar caigui a la segona onada que a la cinquena, posem per cas.

El fet de considerar no tan sols l'ocurrència o no de l'esdeveniment, sinó també el temps en què això succeeix, posa problemes específics a l'anàlisi i, en particular, com afrontar el problema dels casos censurats, és a dir, aquells que no experimenten l'esdeveniment al llarg del període d'observació: quin valor de "temps fins a l'esdeveniment" els hauríem d'assignar?

Els models de la família de les anàlisis de supervivència (o anàlisi d'història dels esdeveniments tal com es coneix en ciències socials[17]) van ser especialment pensats per analitzar el temps fins a l'ocurrència d'un esdeveniment quan la mostra d'estudi es compon d'individus en risc, alguns dels quals experimenten l'esdeveniment i altres no al llarg del període d'observació (casos censurats) (Singer i Willett, 2003, p. 305-324).

Un aspecte important de l'anàlisi de supervivència és la unitat mínima de mesura del temps de què es disposa. Alguns fenòmens només poden ocórrer en moments discrets del temps, com és el cas present, en què l'abandonament del panel només pot ocórrer amb ocasió de cada onada. Altres fenòmens sí que ocorren realment en temps continuu, però només disposem d'una mesura discreta del moment en què van ocórrer. Tant en el primer cas com en el segon, si existeix un gran nombre d'empats (individus que experimenten l'esdeveniment en el mateix moment del temps), és millor utilitzar un model de supervivència a temps discret.

El model de riscos a temps discret

El model de risc a temps discret (*discrete-time hazard model*) (Singer i Willett, 2003); (Rabe-Hesketh i Skrondal, 2012b, pp. 743-796) és un tipus particular d'anàlisi de supervivència adequat quan els esdeveniments no es produeixen en qualsevol moment del temps, sinó que estan limitats a moments específics del mateix.

[16] Per tant no distingim el motiu de l'abandonament. En un 15% dels casos el motiu va ser la impossibilitat de contactar amb la llar, mentre que el 85% restant va rebutjar continuar participant-hi.
[17] Altres expressions emprades per referir-se a aquesta tècnica, amb origen en altres disciplines, són: anàlisi de la durada (*duration analysis*), anàlisi del temps de fallida (*failure time analysis*) i anàlisi de riscos (*hazard analysis*) (Gasch, 2010).

Des del punt de vista estadístic, el model de risc a temps discret que utilitzem aquí és una regressió logística en què allò que es prediu és el logaritme de l'*odds* del risc (*hazard, h*), on risc es defineix com la probabilitat que un individu *i* experimenti l'esdeveniment d'interès en el temps t, condicionada al fet que no l'ha experimentat en el temps t-1.

Es tracta d'un model compost per dos termes: el primer modela el logaritme de l'*odds* del risc com una funció lineal de J indicadors de temps, mentre que el segon hi afegeix l'efecte de P predictors:

$$\text{logit h}(t_{ij}) = [\alpha_1 D_{1ij} + \alpha_2 D_{2ij} + \ldots + \alpha_J D_{Jij}] + [\beta_1 X_{1ij} + \beta_2 X_{2ij} + \ldots + \beta_P X_{Pij}]$$

Mitjançant els J indicadors de temps predim el que s'anomena la "funció de risc basal", que mostra l'efecte pur del temps sobre el risc de pèrdua, sense tenir en compte cap variable predictora.

En realitat, la predicció de la funció de risc basal no té per què fer-se obligatòriament a partir de J variables predictores –en el nostre cas, set variables *dummies* corresponents a cada onada. Aquesta estratègia proporciona el millor ajust possible, però suposa una major complexitat del model, en tant, que fa servir el major nombre de paràmetres possible. Es pot triar modelitzar la funció de risc de manera més parsimoniosa, amb menys paràmetres, al preu d'una pèrdua (relativa) d'ajust.

Les assumpcions del model són les següents:

En tota anàlisi de supervivència, s'assumeix que el mecanisme de censura és no informatiu, és a dir, la censura ocorre de forma independent de l'ocurrència de l'esdeveniment i el grau de risc d'ocurrència del mateix. Per exemple, en un estudi sobre recaiguda en l'alcoholisme després d'un tractament, la censura serà informativa si els individus deixen d'informar a causa del fet que han tornat a beure. En el nostre exemple, la censura és no informativa atès que es produeix només a la fi del període d'observació (Singer i Willett, 2003, p. 318-319). En realitat, ens trobem davant el problema general del possible biaix produït per l'existència de valors perduts, en el context particular de l'anàlisi de supervivència.

Les assumpcions específiques del model presentat són:

Existeix una funció de *log odds* del risc per a cada valor de les variables independents i la seva combinació.

La forma de la funció de *log odds* del risc és idèntica per a tots els valors de les variables independents i la seva combinació, i idèntica a la funció basal.

La distància entre les funcions de *log odds* del risc corresponents a cada valor de les variables independents i la seva combinació és constant al llarg del temps.

En termes completament informals, aquestes assumpcions podrien sintetitzar-se en la idea que l'únic que fan les variables predictores i la seva combinació, és "elevar" la funció basal de *log odds* del risc en una determinada magnitud, sense canviar-ne la forma. La predicció d'aquesta forma, tanmateix, pot fer-se de manera molt flexible, per exemple buscant el màxim ajust a les dades, com és el nostre cas.

Aquestes assumpcions (certament restricitives) sobre el comportament poblacional, naturalment, poden ser o no ratificades per les dades. Existeixen mitjans per comprovar-ne l'acompliment, així com per relaxar aquests supòsits en cas que no es compleixin (Singer i Willett, 2003, p. 366).

3.3.3. Descripció de les dades

Variables d'estudi

A la taula 5 es resumeixen les variables d'estudi.

Taula 5. Variables d'estudi

Tipus	Nom	Descripció	Valors
Dependent o resposta	**Esdeveniment**	Indica si la llar ha deixat de pertànyer a la mostra en la ronda	0: No 1: Sí
Independents o predictores	**Onada**	Número d'onada	1 a 8
	Ruralitat	Combinació del nombre d'habitants del municipi on es troba la llar enquestada amb la seva densitat de població	1: Rural profund 2: Rural 3: Rural dens 4: Urbà
	Idioma	Idioma de preferència de la persona entrevistada	1: Català 2: Castellà 3. Indiferent

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

En primer lloc carreguem el fitxer de dades amb què treballarem en aquest exemple:

```
use "C:\ ATTRITION.DTA", clear                                    (i)
```

Estructura de les dades

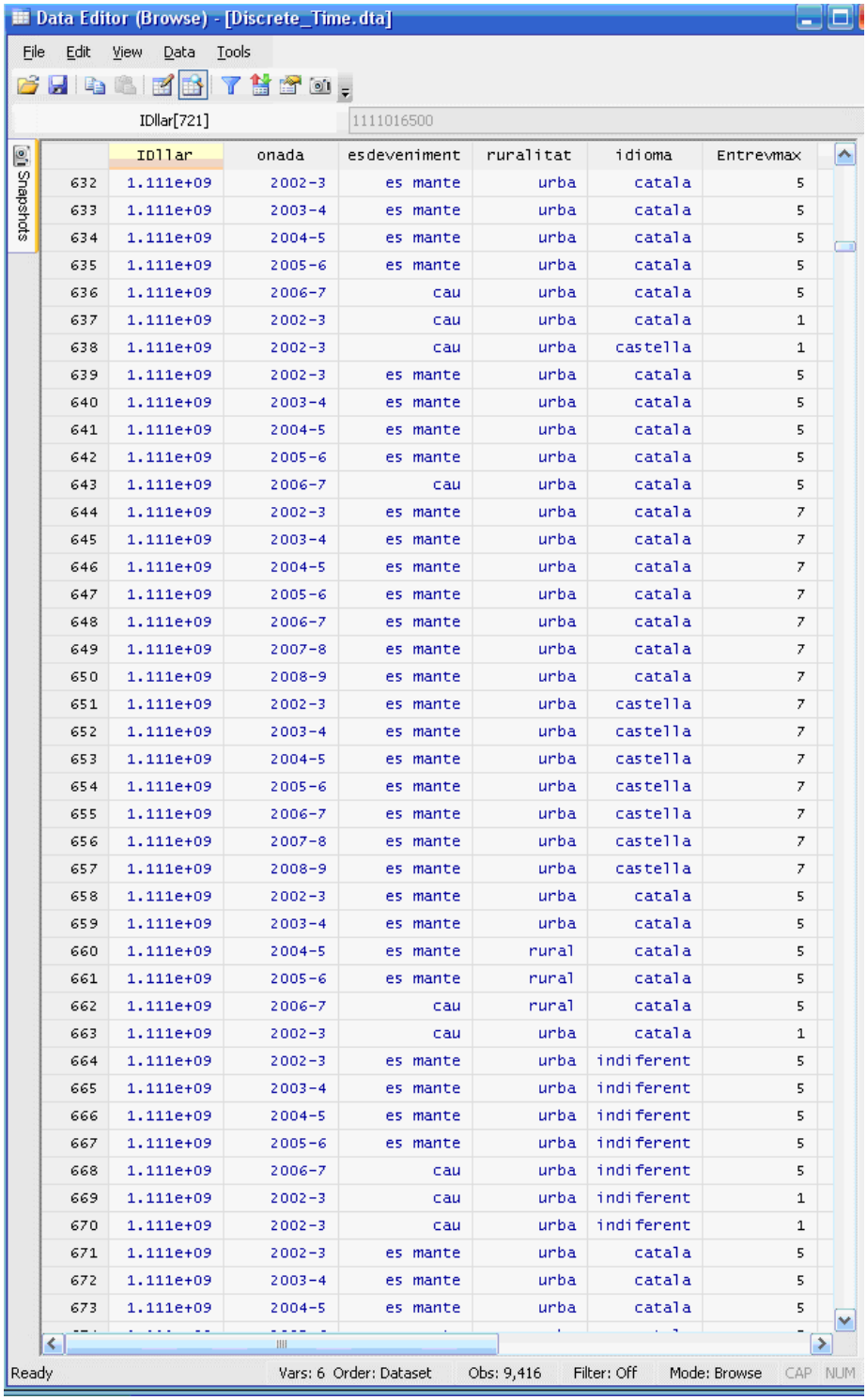
L'anàlisi de supervivència a temps discret ajustat mitjançant regressió logística tal com el duem a terme aquí requereix un format específic de la matriu de dades. Cal que cada llar (IDllar) disposi de registres (files) repetits, tants com rondes hagi participat al panel (més l'onada on es produeix l'abandonament, si s'ha produït). Aquest format es coneix com a "format llarg" (long), en contraposició amb el "format ample" (wide), en què cada llar ocuparia un sol registre i la informació corresponent a les rondes apareix en columnes o variables, una per ronda (vegeu la figura 19).

La variable dependent binària, esdeveniment en aquest cas, val sempre 0 excepte en el darrer registre de la llar, que tindrà el valor 1 si s'ha produït abandonament, o bé 0 si la llar ha continuat al panel fins a la darrera ronda observada (aquesta situació s'anomena "censura", ja que l'esdeveniment no s'ha produït al llarg del període d'observació). En el cas d'aquest estudi, només hi poden haver casos censurats en la darrera onada estudiada (ja que sempre sabem si una llar ha continuat o no al panel) –però no és el cas general ni de bon tros (Singer i Willett, 2003, p. 317).

D'altra banda, atès que en aquest estudi no disposem d'informació sobre les possibles variables predictores corresponents a la ronda en què la llar abandona la mostra (precisament perquè la llar no és entrevistada), els valors de les variables predictores corresponen invariablement a les de l'onada anterior, és a dir: utilitzem els valors de 2002 per predir la pèrdua el 2003, els del 2003 per predir la pèrdua el 2004, etc.

A la figura 19 es mostra l'aspecte de la base de dades, a punt per portar a terme l'anàlisi de supervivència mitjançant el model proposat en aquest estudi de cas.

Figura 19. Matriu de dades



	IDllar	onada	esdeveniment	ruralitat	idioma	Entrevmax
632	1.111e+09	2002-3	es mante	urba	catala	5
633	1.111e+09	2003-4	es mante	urba	catala	5
634	1.111e+09	2004-5	es mante	urba	catala	5
635	1.111e+09	2005-6	es mante	urba	catala	5
636	1.111e+09	2006-7	cau	urba	catala	5
637	1.111e+09	2002-3	cau	urba	catala	1
638	1.111e+09	2002-3	cau	urba	castella	1
639	1.111e+09	2002-3	es mante	urba	catala	5
640	1.111e+09	2003-4	es mante	urba	catala	5
641	1.111e+09	2004-5	es mante	urba	catala	5
642	1.111e+09	2005-6	es mante	urba	catala	5
643	1.111e+09	2006-7	cau	urba	catala	5
644	1.111e+09	2002-3	es mante	urba	catala	7
645	1.111e+09	2003-4	es mante	urba	catala	7
646	1.111e+09	2004-5	es mante	urba	catala	7
647	1.111e+09	2005-6	es mante	urba	catala	7
648	1.111e+09	2006-7	es mante	urba	catala	7
649	1.111e+09	2007-8	es mante	urba	catala	7
650	1.111e+09	2008-9	es mante	urba	catala	7
651	1.111e+09	2002-3	es mante	urba	castella	7
652	1.111e+09	2003-4	es mante	urba	castella	7
653	1.111e+09	2004-5	es mante	urba	castella	7
654	1.111e+09	2005-6	es mante	urba	castella	7
655	1.111e+09	2006-7	es mante	urba	castella	7
656	1.111e+09	2007-8	es mante	urba	castella	7
657	1.111e+09	2008-9	es mante	urba	castella	7
658	1.111e+09	2002-3	es mante	urba	catala	5
659	1.111e+09	2003-4	es mante	urba	catala	5
660	1.111e+09	2004-5	es mante	rural	catala	5
661	1.111e+09	2005-6	es mante	rural	catala	5
662	1.111e+09	2006-7	cau	rural	catala	5
663	1.111e+09	2002-3	cau	urba	catala	1
664	1.111e+09	2002-3	es mante	urba	indiferent	5
665	1.111e+09	2003-4	es mante	urba	indiferent	5
666	1.111e+09	2004-5	es mante	urba	indiferent	5
667	1.111e+09	2005-6	es mante	urba	indiferent	5
668	1.111e+09	2006-7	cau	urba	indiferent	5
669	1.111e+09	2002-3	cau	urba	indiferent	1
670	1.111e+09	2002-3	cau	urba	indiferent	1
671	1.111e+09	2002-3	es mante	urba	catala	5
672	1.111e+09	2003-4	es mante	urba	catala	5
673	1.111e+09	2004-5	es mante	urba	catala	5

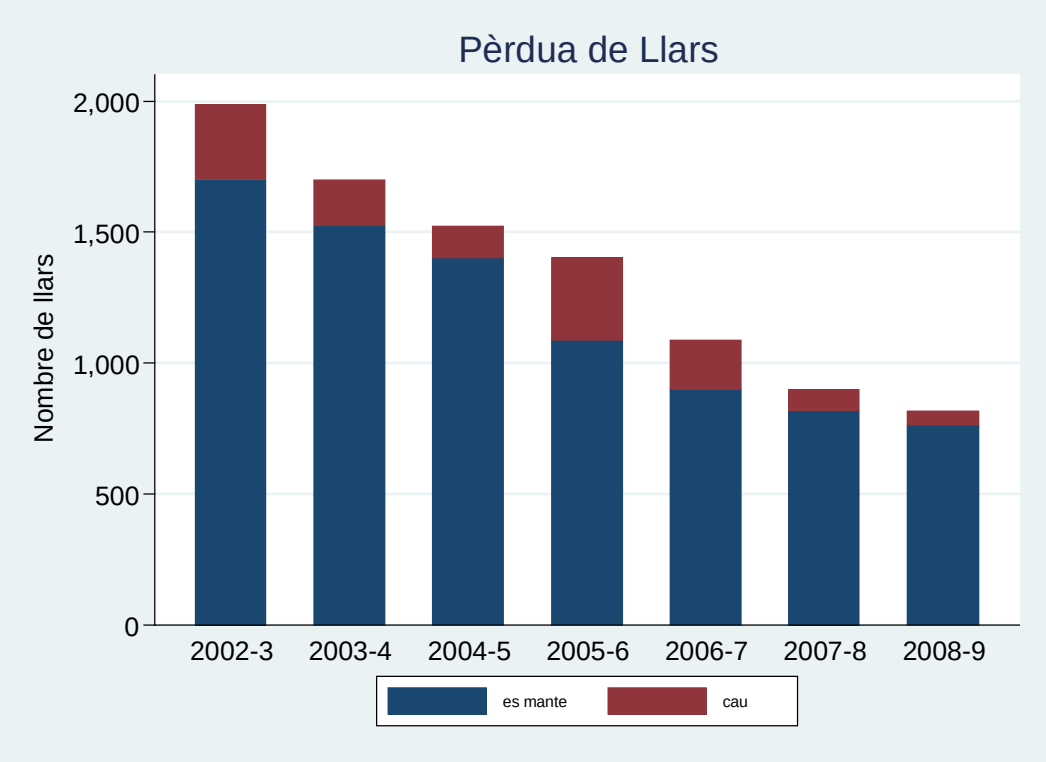
Font: elaboració pròpia a partir de dades del Panel de Desigualtats Socials a Catalunya, 2002-2009.

La pèrduda de mostra o attrition en el PaD en el període 2002-2009

Podem començar elaborant un gràfic que descrigui l'evolució del desgast mostral (vegeu la figura 20) mitjançant el comandament següent (`ii`):

```
graph bar (count) IDllar, over(esdeveniment) over(onada) asyvars stack  (ii)
legend(cols(3) size(vsmall))ytitle(Nombre de llars) ylabel(, angle(0)
format(%12.0gc)) title(Attrition de Llar)
```

Figura 20. Evolució del desgast de la mostra de llars del Panel de Desigualtats (PaD). Catalunya, 2002-2009



Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

El gràfic fa palès que el desgast mostral és molt considerable al llarg del període i que la major pèrdua en termes absoluts es produeix en les rondes corresponents al 2003 (la que segueix a la inicial) i al 2006.

La taula de vida

En un estudi de supervivència, una manera habitual de descriure les dades és mitjançant la taula de vida (*life table*), que permet observar amb detall la distribució del desgast al llarg del temps i obtenir una funció basal de risc i de supervivència.

En Stata, el comandament per generar taules de vida és `ltable`. A diferència de la regressió logística que ajustarem després, aquest comandament requereix per executar-se correctament un únic registre per llar, el corresponent a la seva darrera participació en el panel. Per tant, fem una selecció temporal d'aquests registres mitjançant els comandaments `preserve-restore`) abans d'executar `ltable(iii)`.

```
preserve                                                                (iii)
keep if onada-2001==Entrevmax
ltable onada esdeveniment, survival hazard noadjust
```

El quadre 26 mostra els resultats. En les tres primeres columnes de la primera part es mostren els

efectius en risc a l'inici de cada període, les caigudes del panel i els casos censurats (*lost*). La resta de columnes mostren la funció de supervivència i els seus intervals de confiança.

La segona part de la taula mostra en primer lloc el complementari a 1 de la funció de supervivència, i en segon lloc la funció de risc, acompanyades dels seus intervals de confiança respectius.

En el període 2002-2009 el PaD va sofrir un desgast important de la mostra: 1.225 llars de les 1.987 de la mostra original del 2002, és a dir el 61,7% van abandonar el panel. La major pèrdua de llars en termes absoluts va tenir lloc el 2006, amb 314 llars perdudes. La segona pèrdua en importància correspon a la segona ronda (2003), quan van caure 288 llars.

La probabilitat condicionada (risc o *hazard*) de caure del panel en la ronda següent és màxima en les rondes del 2005 i el 2006 (0.22; IC95%: 0.20; 0.25) i entre les del 2006 i el 2007: 0.17 (IC95%: 0.15; 0.20).

Quadre 26. Taula de vida. PaD, Catalunya, 2002-2009

Interval		Beg. Total	Deaths	Lost	Survival	Std. Error	[95% Conf.	Int.]
2002	2003	1987	288	0	0.8551	0.0079	0.8388	0.8698
2003	2004	1699	175	0	0.7670	0.0095	0.7478	0.7850
2004	2005	1524	123	0	0.7051	0.0102	0.6845	0.7246
2005	2006	1401	314	0	0.5471	0.0112	0.5249	0.5686
2006	2007	1087	187	0	0.4529	0.0112	0.4309	0.4747
2007	2008	900	82	0	0.4117	0.0110	0.3900	0.4332
2008	2009	818	56	762	0.3835	0.0109	0.3621	0.4048

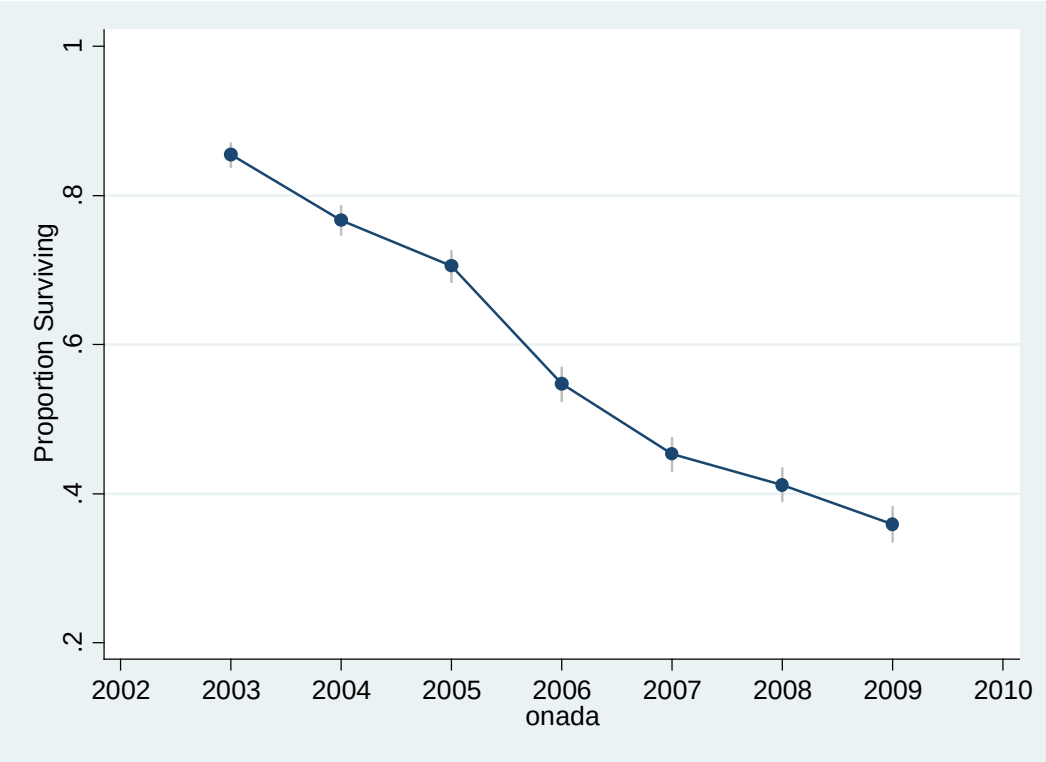
Interval		Beg. Total	Cum. Failure	Std. Error	Hazard	Std. Error	[95% Conf.	Int.]
2002	2003	1987	0.1449	0.0079	0.1449	0.0085	0.1287	0.1622
2003	2004	1699	0.2330	0.0095	0.1030	0.0078	0.0883	0.1188
2004	2005	1524	0.2949	0.0102	0.0807	0.0073	0.0671	0.0956
2005	2006	1401	0.4529	0.0112	0.2241	0.0126	0.2000	0.2496
2006	2007	1087	0.5471	0.0112	0.1720	0.0126	0.1483	0.1975
2007	2008	900	0.5883	0.0110	0.0911	0.0101	0.0725	0.1119
2008	2009	818	0.6165	0.0109	0.0685	0.0091	0.0517	0.0875

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

Un gràfic de la funció de supervivència (figura 21) i els seus intervals de confiança pot obtenir-se mitjançant el comandament `(iv)`:

```
ltable onada esdeveniment, survival graph ci notable xlabel(2002 (1) (iv)
2010)
```

Figura 21. Gràfic de la funció de supervivència



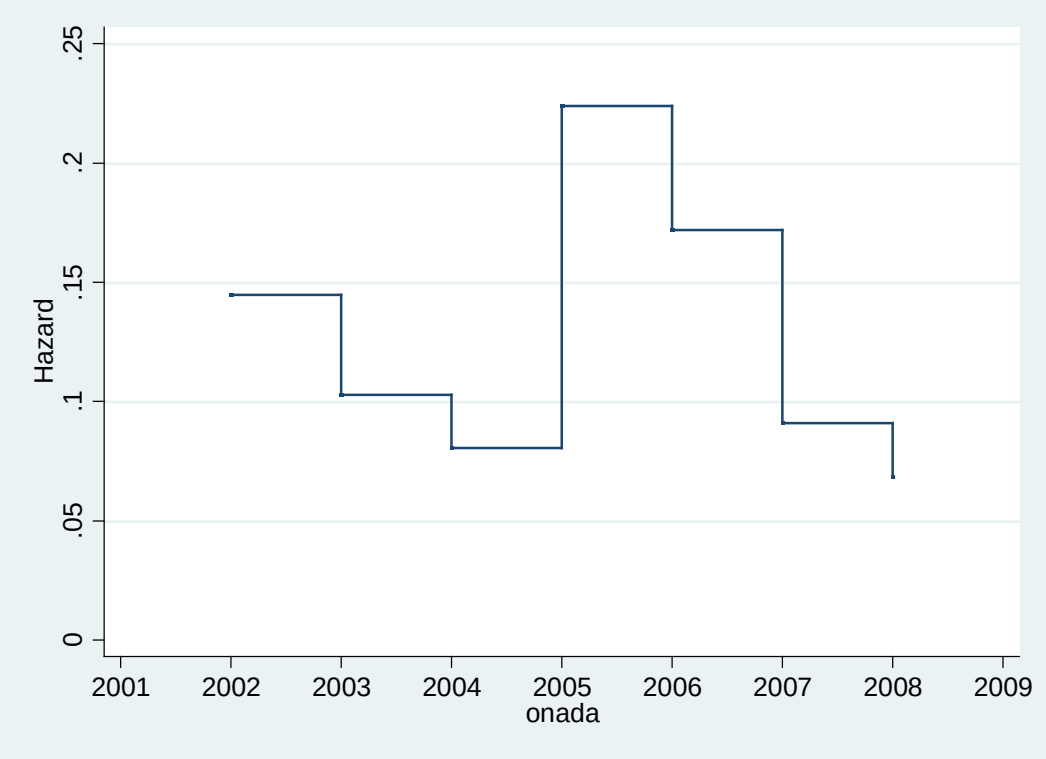
Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

Si en canvi allò que volem mostrar en el gràfic és el risc d'abandonament de la mostra (figura 21) podem fer servir el comandament següent `(v)`:

```
stset onada2, id(IDllar) failure(esdeveniment)                    (v)
sts gen h=h

graph twoway scatter h onada, msymbol(i) connect(stairstep) yla-
bel(0(.05).25) xlabel(2001(1)2009) xtitle(Onada/Any) ytitle(Hazard)
sort

restore
```

La forma inusual de la funció de risc (habitualment en els estudis de panel el risc és màxim després de la primera onada i tendeix a disminuir) és consistent amb el fet que el panel estava previst inicialment només per a cinc rondes. L'intent d'ampliar el període inicial de compromís de les llars fou rebutjat per moltes.

Figura 22. Funció de risc de la pèrdua de mostra de llars al Panel de Desigualtats a Catalunya, 2002-2009



Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

3.3.4. Execució, interpretació de resultats i comentaris amb Stata

Ajust del model de riscos a temps discret

Per a analitzar el risc de pèrdua de les llars i determinar els possibles predictors s'ajustaran tres models. En primer lloc un model de risc basal, sense cap altra variable predictora que el temps, determinat pel número d'onada (model 1, quadre 27). Seguidament hi afegim les variables "ruralitat" (model 2.1, quadre 28) i "idioma" (model 2.2, quadre 29) per separat, per comprovar-ne la significació. Finalment, en un tercer model s'hi inclouen els dos predictors alhora (model 3, quadre 30).

Prèviament incrementem en una unitat el valor d'onada, per fer que les seves categories siguin més fàcilment interpretables:

```
replace onada=onada+1                                                    (vi)
```

El comandament per ajustar el model 1 (quadre 27) és (vii):

```
logit esdeveniment bn.onada, nocons or                    (vii)
```

En afegir el terme `nocons` demanem que el model s'ajusti sense constant; en cas contrari, no obtindríem estimació per a la primera onada de risc (2003). L'opció `or` en aquest cas és enganyosa, ja que la columna corresponent no ens ofereix *odds ratios* (OR), sinó simplement *odds* (O) (precisament perquè manca un any de referència). Els resultats obtinguts són equivalents als de la funció de risc de la taula de vida, ja que $h(t) = \frac{O_{(t)}}{1+O_{(t)}}$ p.e. $\frac{0.1695}{1+0.1695} = 0.1449$, que és el risc de caiguda del panel per a l'any 2003.

NOTA: en el paquet estadístic Stata v.12, les variables *dummy* (en aquest cas les corresponents a "onada") es generen automàticament afegint davant la variable d'interès el prefix bn. (si volem *dummies* per a totes les categories *k*) o el prefix *i.* (si volem *dummies* per a k-1 categories, deixant la primera de referència).

Quadre 27. Model 1. Risc basal de pèrdua

```
Iteration 0:   log likelihood = -6526.6739
Iteration 1:   log likelihood = -3555.6098
Iteration 2:   log likelihood = -3536.6164
Iteration 3:   log likelihood =  -3536.478
Iteration 4:   log likelihood =  -3536.478

Logistic regression                               Number of obs   =       9416
                                                  Wald chi2(7)    =    3687.99
Log likelihood =  -3536.478                       Prob > chi2     =     0.0000
```

esdeveniment	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada						
2003	.1695115	.010802	-27.85	0.000	.1496087	.192062
2004	.1148294	.0091651	-27.12	0.000	.0982006	.134274
2005	.0877944	.0082563	-25.87	0.000	.0730161	.1055639
2006	.2888684	.0185072	-19.38	0.000	.2547801	.3275176
2007	.2077778	.0166983	-19.55	0.000	.1774971	.2432242
2008	.1002445	.0116118	-19.86	0.000	.0798844	.1257937
2009	.0734908	.0101751	-18.86	0.000	.0560249	.0964018

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

Amb els comandaments `(viii)` i `(ix)` explorem la relació entre la pèrdua de llars i les variables potencialment predictores "ruralitat" (model 2.1) i "idioma" (model 2.2), respectivament.

```
logit esdeveniment bn.onada i.ruralita, nocons or         (viii)
```

Quadre 28. Model 2.1. Efecte cru de la dimensió rural-urbà sobre el desgast mostral

```
Iteration 0:   log likelihood = -6526.6739
Iteration 1:   log likelihood = -3539.8759
Iteration 2:   log likelihood = -3516.4438
Iteration 3:   log likelihood = -3516.1935
Iteration 4:   log likelihood = -3516.1932

Logistic regression                               Number of obs   =       9416
                                                  Wald chi2(10)   =    3649.77
Log likelihood = -3516.1932                       Prob > chi2     =     0.0000
```

esdeveniment	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada						
2003	.0589477	.0155942	-10.70	0.000	.0350984	.0990026
2004	.0399966	.0107562	-11.97	0.000	.0236108	.0677541
2005	.030678	.008381	-12.75	0.000	.0179591	.0524045
2006	.1018448	.0268463	-8.67	0.000	.0607522	.1707321
2007	.0734236	.0196393	-9.76	0.000	.0434667	.1240264
2008	.0354166	.0099272	-11.92	0.000	.0204465	.0613474
2009	.0259579	.0075393	-12.57	0.000	.0146907	.0458664
ruralitat						
2	2.248889	.6056802	3.01	0.003	1.326535	3.812564
3	2.784941	.750159	3.80	0.000	1.642608	4.721694
4	3.177876	.8284535	4.44	0.000	1.906487	5.297124

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

En relació amb la categoria de referència (1 "rural profund"), la resta de categories presenten un risc significativament major de caiguda del panel, particularment la categoria 4, corresponent a l'hàbitat urbà (OR: 3.18; IC95%: 1.91; 5.30).

```
logit esdeveniment bn.onada i.idioma, nocons or                    (ix)
```

Quadre 29. Model 2.2. Efecte cru de l'idioma de preferència per a l'entrevista sobre el desgast mostral

```
Iteration 0:   log likelihood = -6526.6739
Iteration 1:   log likelihood = -3532.0502
Iteration 2:   log likelihood = -3510.3669
Iteration 3:   log likelihood = -3510.2038
Iteration 4:   log likelihood = -3510.2038
```

```
Logistic regression                                     Number of obs   =       9416
                                                        Wald chi2(9)    =    3654.02
Log likelihood = -3510.2038                             Prob > chi2     =     0.0000
```

esdeveniment	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada						
2003	.1421874	.0098405	-28.18	0.000	.1241513	.1628437
2004	.0972098	.0081637	-27.76	0.000	.0824567	.1146024
2005	.0750227	.0073013	-26.61	0.000	.0619945	.0907888
2006	.2505132	.0169687	-20.44	0.000	.2193683	.28608
2007	.1803097	.0150458	-20.53	0.000	.1531057	.2123474
2008	.0872559	.0103007	-20.66	0.000	.0692323	.1099715
2009	.0640037	.0089834	-19.58	0.000	.0486109	.0842709
idioma						
2	1.630313	.1108697	7.19	0.000	1.426872	1.862761
3	1.476709	.2152805	2.67	0.007	1.109694	1.965108

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

En relació amb la categoria de referència (1 "català"), l'elecció de "castellà" per a la realització de l'entrevista presenta un risc significativament major de caiguda del panel (OR: 1.48; IC95%: 1.11; 1.97). L'opció "indiferent" presenta igualment major risc de caiguda.

Finalment, veiem l'efecte dels dos predictors, "ruralitat" i "idioma", tots junts (quadre 30) mitjançant la instrucció `(x)`:

```
logit esdeveniment bn.onada i.ruralita i.idioma, nocons or                (x)
```

Quadre 30. Model 3. Efecte de l'eix rural-urbà i l'idioma de preferència per a l'entrevista sobre el desgast mostral

```
Iteration 0:   log likelihood = -6526.6739
Iteration 1:   log likelihood = -3524.1517
Iteration 2:   log likelihood = -3499.2907
Iteration 3:   log likelihood =  -3499.026
Iteration 4:   log likelihood = -3499.0257

Logistic regression                                     Number of obs   =       9416
                                                        Wald chi2(12)   =    3629.27
Log likelihood = -3499.0257                             Prob > chi2     =     0.0000
```

esdeveniment	Odds Ratio	Std. Err.	z	P>|z|	[95% Conf.	Interval]
onada						
2003	.0578653	.0153138	-10.77	0.000	.034447	.097204
2004	.0395465	.0106393	-12.01	0.000	.0233404	.0670053
2005	.0305291	.0083429	-12.77	0.000	.017869	.0521589
2006	.1022965	.0269697	-8.65	0.000	.0610166	.1715037
2007	.0738456	.0197557	-9.74	0.000	.0437125	.1247509
2008	.0357351	.0100184	-11.88	0.000	.0206281	.0619059
2009	.026232	.0076203	-12.53	0.000	.0148443	.0463558
ruralitat						
2	2.164581	.5832772	2.87	0.004	1.276455	3.670643
3	2.592637	.6995328	3.53	0.000	1.527828	4.399555
4	2.680613	.703869	3.76	0.000	1.602241	4.484774
idioma						
2	1.513435	.108133	5.80	0.000	1.315669	1.740929
3	1.374342	.2018557	2.16	0.030	1.030565	1.832798

Font: Panel de Desigualtats Socials a Catalunya, 2002-2009.

En aquest model multivariat, els efectes "crus" no es veuen substancialment afectats: les OR continuen tenint valors similars. Podem concloure doncs que tant el grau d'urbanització del territori com l'idioma de realització de l'entrevista afecten significativament el risc de les llars de caure del panel.

L'anàlisi duta a terme aquí podria estendre's en diverses direccions: per exemple dintre de l'esdeveniment d'interès podríem considerar interessant distingir entre pèrdua deguda a refús, o pèrduda deguda a no-localització de la llar, i comprovar si les variables associades a cada tipus d'esdeveniment són les mateixes; o bé podríem considerar que existeix heterogeneïtat no observada entre les llars, o dit d'una altra manera, que les llars difereixen de forma significativa en característiques no observades (*frailty models*), la qual cosa ens duria a incloure efectes aleatoris en el model de supervivència.

4. Referències bibliogràfiques

ABBOTT, A.; FORREST, J. (1986). "Optimal Matching Methods for Historical Sequences". *Journal of Interdisciplinary History*, núm. 16(3), p. 471. DOI: *http://dx.doi.org/10.2307/204500*

ALLISON, P. D. (2001). *Missing data*. Thousand Oaks, Calif.: Sage Publications.

ALLISON, P. D. (2005). *Fixed Effects Regression Methods For Longitudinal Data Using Sas*. SAS Publishing.

ANDRUFF, H.; CARRARO, N.; THOMPSON, A.; GAUDREAU, P.; LOUVET, B. (2009). "Latent Class Growth Modelling: A Tutorial". Disponible a: *http://core.kmi.open.ac.uk/display/896472*

BAUM, C. F. (2006). *An Introduction to modern econometrics using Stata*. College Station: Stata Press.

BEHR, A.; BELLGARDT, E.; RENDTEL, U. (2005). "Extent and Determinants of Panel Attrition in the European Community Household Panel". *European Sociological Review*, núm. 21(5), p. 489-512. DOI: *http://dx.doi.org/10.1093/esr/jci037*

BERGMAN, L R I MAGNUSSON, D. (1997). "A person-oriented approach in research on developmental psychopathology". *Development and Psychopathology*, núm. 9(2), p. 291-319.

BERGMAN, LARS R.; MAGNUSSON, D.; EL-KHOURI, B. (2003). *Studying individual development in an interindividual context: a person-oriented approach*. Mahwah, N.J.: L. Erlbaum Associates.

BERGSMA, W. P.; CROON, MARCEL A.; HAGENAARS, J. A. (2009). *Marginal models for dependent, clustered, and longitudinal categorical data*. Nova York – Londres: Springer. Disponible a: *http://www.myilibrary.com?id=206868*

BLANCHARD, P.; BUHLMANN, F.; GAUTHIER, J.-A. (2012). "Sequence Analysis in 2012". Presentat a Conference on Sequence Analysis (LaCOSA), Lausanne. Disponible a: *http://www3.unil.ch/wpmu/sequences2012/files/2012/06/LaCOSA_Introduction.pdf*

CAMERON, A. C.; TRIVEDI, P. K. (2010). *Microeconometrics using Stata*. College Station, Tex.: Stata Press.

COLLINS, L. M. (2006). "Analysis of longitudinal data: the integration of theoretical model, temporal design, and statistical model". *Annual Review of Psychology*, 57, p. 505-528. DOI: *http://dx.doi.org/10.1146/annurev.psych.57.102904.190146*

COLLINS, L. M.; LANZA, S. T. (2010). *Latent Class and Latent Transition Analysis: With Applications in the Social, Behavioral, and Health Sciences*. John Wiley & Sons.

ENDERS, C. K. (2010). *Applied Missing Data Analysis*. Guilford Press.

FINKEL, S. E. (1995). *Causal analysis with panel data*. Thousand Oaks, Calif.: Sage Publications.

FITZMAURICE, G. M.; DAVIDIAN, M.; VERBEKE, G.; MOLENBERGHS, G. (2009). *Longitudinal data analysis*. Chapman & Hall/CRC.

FITZMAURICE, G. M.; MOLENBERGHS, G. (2009). "Advances in longitudinal data analysis: An historical perspective. Dins: FITZMAURICE, G. M.; DAVIDIAN, M.; VERBEKE, G.; MOLENBERGHS, G. (2009). *Longitudinal data analysis,* p. 3-27. Chapman & Hall/CRC.

FUNDACIÓ DEL MÓN RURAL; ALDOMÀ BUIXADÉ, I. (2009). "Socialització de la variable rural-urbà 1a a 11a onades del PaD". Document no publicat.

FUNDACIÓ JAUME BOFILL (2009). "Enquesta Panel de Desigualtats a Catalunya (Aspectes Metodològics)".

GABADINHO A.; RITSCHARD, G.; STUDER, M.; MÜLLER, N. S. (2011). *Mining sequence data in R with the TraMineR package: A user's guide*. Ginebra: University of Geneva.

GABADINHO ALEXIS, R. G. (2011). "Analyzing and Visualizing State Sequences in R with TraMineR". *Journal of Statistical Software*.

GAUTHIER CÉDRIC NOTREDAME, J.-A. (2009). "How Much Does It Cost? Optimization of Costs in Sequence Analysis of Social Science Dat"a. *Sociological Methods Research*, núm. 38(1), p. 197-231.

GREENACRE, M. (2008). *La práctica del análisis de correspondencias*. Bilbao: Fundación BBVA.

GREENACRE, M. J.; BLASIUS, J. (1994). *Correspondence analysis in the social sciences: recent deve-*

lopments and applications. Londres – San Diego, CA: Academic Press.

HALABY, C. N. (2004). "Panel Models in Sociological Research: Theory into Practice". *Annual Review of Sociology*, núm. 30(1), p. 507-544. DOI: *http://dx.doi.org/10.1146/annurev.soc.30.012703.110629*

JAVIER, C.; BART, G. (s.d.). "Appraisal of several methods to model time to multiple events per subject: modelling time to hospitalizations and death". *Revista Colombiana de Estadística*, núm. 33(1), p. 43-61.

JIMENEZ-MARTIN, S.; PERACCHI, F. (2002). "Sample attrition and labor force dynamics: Evidence from the spanish labor force survey". *Spanish Economic Review*, núm. 4(2), p. 79-102.

KAUFMAN, L.; ROUSSEEUW, P. J. (2005). *Finding groups in data*. Hoboken: John Wiley & Sons.
LEE, U. (2003). "Panel Attrition in Survey Data: A Literature Review". Disponible a: *http://www.cssr.uct.ac.za/sites/cssr.uct.ac.za/files/pubs/wp41.pdf*

LEMAY, M. (2009). *Understanding the Mechanism of Panel Attrition*. University of Maryland. Disponible a: *http://hdl.handle.net/1903/9631*

MAGNUSSON, D.; BERGMAN, L. R.; RUDINGER, G. (1994). *Problems and Methods in Longitudinal Research: Stability and Change*. Cambridge: Cambridge University Press.

MALLAPRAGADA, P. K.; JIN, R.; JAIN, A. (2010). "Non-parametric mixture models for clustering". *Proceedings of the 2010 joint IAPR international conference on Structural, syntactic, and statistical pattern recognition*, p. 334-343.

MASSONI, S.; OLTEANU, M.; ROUSSET, P. (2009). "Career-path analysis using optimal matching and self-organizing maps". A *Advances in Self-Organizing Maps* (p. 154-162). Springer. Disponible a: *http://link.springer.com/chapter/10.1007/978-3-642-02397-2_18*

MENARD, S. W. (2002). *Longitudinal research*. Thousand Oaks, Calif.: Sage Publications.

MOLENBERGHS, G.; VERBEKE, G. (2005). *Models for discrete longitudinal data*. Springer.

PERACCHI, F. (2002). "The European Community Household Panel: A review". *Empirical Economics Empirical Economics*, núm. 27(1), p. 63-90.

RABE-HESKETH, S.; SKRONDAL, A. (2012a). *Multilevel and Longitudinal Modeling Using Stata. Volume I: Continuous responses*. 3a ed., Vol. I. Stata Press Publication.

RABE-HESKETH, S.; SKRONDAL, A. (2012b). *Multilevel and Longitudinal Modeling Using Stata. Volume II*. 3a ed., vol. II. Stata Press Publication.

RIVERO RODRÍGUEZ, G. (2011). *Análisis de datos incompletos en Ciencias Sociales*. Madrid: CIS.

SERVEI D'ESTADÍSTICA UAB (2012). "Projecte Imputació Ingressos Anuals a Nivell Individual Onades B-H del Panel de Desigualtats Socials a Catalunya". Document inèdit.

SINGER, J. D.; WILLETT, J. B. (2003). *Applied longitudinal data analysis: modeling change and event occurrence*. Oxford: Oxford University Press.

SKRONDAL, A.; RABE-HESKETH, S. (2013). "Handling initial conditions and endogenous covariates in dynamic/transition models for binary data with unobserved heterogeneity". *Journal of the Royal*

Statistical Society: Series C (Applied Statistics), n/a-n/a. DOI: *http://dx.doi.org/10.1111/rssc.12023*

STERBA, S. K. I BAUER, D. J. (2010). "Matching method with theory in person-oriented developmental psychopathology research". *Development and Psychopathology*, núm. 22(02), p. 239. DOI: *http://dx.doi.org/10.1017/S0954579410000015*

TARIS, T. (2000). *A primer in longitudinal data analysis*. SAGE.

VAN BERG, G. J. D.; LINDEBOOM, M.; RIDDER, G. (1994). "Attrition in longitudinal panel data and the empirical analysis of dynamic labour market behaviour". *Journal of Applied Econometrics*, núm. 9(4), p. 421-435. DOI: *http://dx.doi.org/10.1002/jae.3950090408*

VERD, J. M.; LÓPEZ-ANDREU, M. (2012). "La inestabilidad del empleo en las trayectorias laborales. Un análisis cuantitativo". *Revista Española de Investigaciones Sociológicas*, núm. 138, p. 135-148. DOI: *http://dx.doi.org/10.5477/cis/reis.138.135*

ZUBIRI-REY, J. B. (2008). *Trayectorias sociolaborales: introducción metodológica a las técnicas longitudinales en economía del trabajo* (Post-Print). HAL. Disponible a: *http://econpapers.repec.org/paper/haljournl/halshs-00371856.htm*