Methods for Longitudinal Data

Gilbert Ritschard

Institute for demographic and life course studies, University Geneva
http://mephisto.unige.ch

Doctoral Program, Lausanne, May 20, 2011

The IP 14 Team

- This day is brought to you by the IP 14 team:
  - Gilbert Ritschard, IDEMO, UNIGE
  - Paolo Ghisletta, FPSE, UNIGE
  - André Berchtold, IMA, UNIL
  - Reto Schumacher, IDEMO, UNIGE
  - Jacques-Antoine Gauthier, MISC, UNIL
  - Delphine Courvoisier, HUG and FPSE, UNIGE
- Associated researchers and PhD students:
  - Alexis Gabadinho
  - Danilo Bolano
  - Reto Bürgin
  - Emmanuel Rousseaux
- TraMineR team:
  - Alexis Gabadinho, Nicolas S. Müller and Matthias Studer

Objective of this doctoral school day

- Provide an introduction to quantitative methods for life course analysis
- Overview of the various longitudinal analysis approaches
- At the end of this day, you should be able to
  - distinguish between various types of longitudinal data;
  - recognize different ways of organizing longitudinal data;
  - identify questions and issues related to longitudinal data;
  - select an appropriate method for the research question and data at hand.
- Insight to IP 14: "Measuring life sequences and the disorder of lives"

What you will not learn today

- Practice of the methods
- Details about the methods
- Expertise in specialized softwares

Outline

Life course

- Life course: "a sequence of socially defined events and roles that the individual enacts over time" (Giele and Elder, 1998, p22)
- Focus on the connection between individuals and the historical and socioeconomic context in which these individuals lived (Elder, 1974)
- Studying life course means tracking individual trajectories (micro level)
- as opposed to macro analysis that follow aggregates (number of divorces, unemployment rates, ...) over time
- Longitudinal data is the fundamental material for empirical analysis of the life course

What is longitudinal data?

Longitudinal data

- Repeated observations on units observed over time (Beck and Katz, 1995).
- "A dataset is longitudinal if it tracks the same type of information on the same subjects at multiple points in time". (http://www.caldercenter.org/whatis.cfm)
- "The defining feature of longitudinal data is that the multiple observations within subject can be ordered" (Singer and Willett, 2003)

Successive transversal data vs longitudinal data

- Successive transversal observations (same units)

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

- Longitudinal observations

id	t_1	t_2	t_3	$\cdots$
1	B	B	D	$\cdots$
2	A	B	C	$\cdots$
3	B	B	A	$\cdots$

Repeated independent cross sectional observations

- Successive independent transversal observations

id	t_1	t_2	t_3	...
11	B	.	.	...
12	A	.	.	...
13	B	.	.	...
.	.	.	.	...
21	.	B	.	...
22	.	B	.	...
23	.	B	.	...
.	.	.	.	...
24	.	.	D	...
25	.	.	C	...
26	.	.	A	...
.	.	.	.	...

- This is not longitudinal ...
- but ... sequences of transversal (aggregated) characteristics.

Longitudinal data: Where do they come from?

- Individual follow-ups: Each important event is recorded as soon as it occurs (medical card, cellular phone, ...).
- Panels: Periodic observation of same units
- Retrospective data (biography): Depends on interviewees' memory
- Matching data from different sources (successive censuses, tax data, social security, population registers, acts of marriages, acts of deaths, ...)

  Examples: Wanner and Delaporte (2001), censuses and population registers, Perroux and Oris (2005), 19th Century Geneva, censuses, acts of marriage, registers of deaths, register of migrations.

- Rotating panels: partial follow up

  e.g.; Swiss Labor Force Survey, SLFS, 5 year-rotating panel (Wernli, 2010)

Types of longitudinal data
Numerical vs categorical

- What are we observing (measuring) over time?
- A numerical (scale) variable
  - continuous (many different values): intellectual ability, perceived health level, confidence in political authorities, ...
  - discrete (few different values): number of childbirths, family size, ...
- A categorical variable
  - States: marital status, occupational status, living arrangement, ...
  - Events: divorce, loss of job, contracting illness, ...
- Specific methods for each type of data

Types of longitudinal data
Repeated measures vs time stamped observations

- Repeated measure: list of successive values (panels).

  Time stamp associated to the position in sequence
  - sequence of numeric values

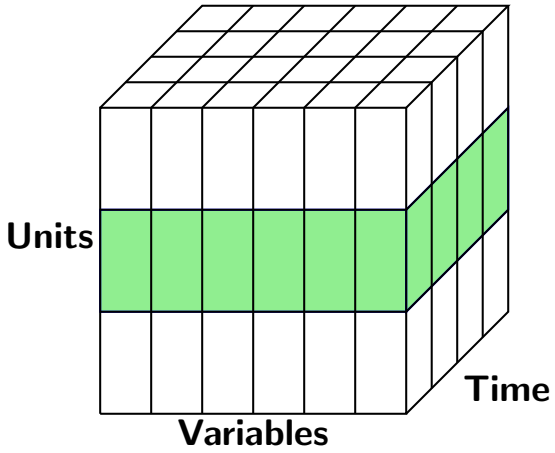
    e.g.; auto-evaluated health, financial situation (..., 10, 10, 8, 7, 7, 5, ...)
  - sequence of states

    (..., married, married, married, divorced, divorced, ...)
- Time stamped events (retrospective surveys)

  (ending school at 17, first job at 17, first union at 20, childbirth at 23, ...)

Organizing longitudinal data

How can we organize three dimensions into row-column form ?



Organizing panel data

- Person level: Put the successive cross-sectional tables next to each other (horizontal organization)
  - Columns can be grouped by variables (instead of times): sequences
- Person-period form: Put the cross-sectional tables above each other (vertical organization)
  - Rows can indeed be sorted by units (instead of time)
- There are plenty of variations which consist essentially in compacting the representation by
  - providing start and end time in a given state
    e.g.: (start, end, state)
    instead of the state or value of the variable at each time point.
    (Ritschard et al., 2009)

Data organization: example for a single unit

Time stamped event

ending secondary school in 1970	first job in 1971	marriage in 1973

State sequences

year	1969	1970	1971	1972	1973
marital status	single	single	single	single	married
education level	primary	secondary	secondary	secondary	secondary
job	no	no	1st	1st	1st

Episodes

id	from	to	marital status	education	emploi
id1	1969	1969	single	primary	no
id1	1970	1970	single	secondary	no
id1	1971	1972	single	secondary	1st
id1	1973	1973	married	secondary	1st

Data organization (1)

- Person level. One row per individual (time in months)

	Job					marriage					...
indiv	beg 1	end 1	beg 2	end 2	...	beg 1	end 1	beg 2	end 2	...	...
1	204	216	260	350	...	300	-	-	-	...	...
2	240	400	401	-	...	340	500	-	-	...	...
⋮											

Data organization: Example of person-episode data

- Person-episode data. One row per spell (new row after any change; i.e.; each time a new event occurs)

indiv	start	end	job	nbre of jobs	married	...
1	1	203	no	0	no	...
1	204	216	yes	1	no	...
1	217	259	no	1	no	...
1	260	299	yes	2	no	...
1	300	350	yes	2	yes	...
2	1	239	no	0	no	...
2	240	339	yes	1	no	...
2	340	400	yes	1	yes	...
2	401	500	yes	2	yes	...
2	...					
⋮						

Data organization: Example of person-period data

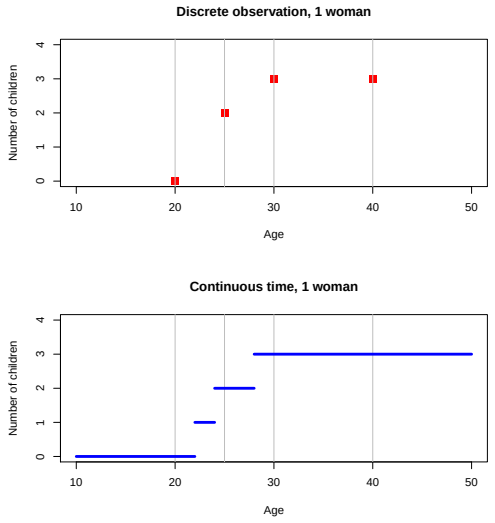
- Person-period form. For each individual, a row for each time unit (day, month, year, ...) where it is observed.

case	month	job	nbr employees	married	...
1	1	no	0	no	...
1	2	no	0	no	...
⋮					
1	350	yes	2	yes	...
2	1	no	0	no	...
⋮					
2	500	yes	2	yes	...
3	1	no	0	no	...
⋮					

Discrete versus continuous time

- Panel data: observations at successive discrete time points (year)
- In other cases (e.g.; retrospective survey) events (such as start of a spell in a given state) may be time stamped using a (almost-)continuous time scale.
  - She got married on June 23, 2004 and had a childbirth on January 15, 2004.
- Sometimes, we may also just be interested in the sequencing (i.e., the order in which event or state occur) and, hence, use an ordinal time scale.
  - Did the childbirth occur before or after marriage?
- Which time granularity? Year? Month? Day?

Discrete vs continuous follow-up



Calendar versus process time (clocks)

- An often important question in longitudinal analysis is time alignment.
  - For instance, aligning on age, we can say that by leaving home at 30 and getting married at 35, she had a late transition to adulthood.
- Panel data are most often aligned on a calendar axis (i.e., the successive observation times)
- Process time such as age, time since marriage, time since end of education, ... are often more interesting from a life course perspective.
- Switching between calendar and process time gives rise to missing data

Which method for which analysis?

- Two types of methods:
  - Exploratory (descriptive) methods:
    - Understand the data and gain knowledge on its intrinsic structure
    - Assumption free, data driven
  - Causal (explanatory, confirmatory) methods
    - Validate research hypotheses
    - Test effects of covariates (sex, birth cohorts, ...) on a target variable (trajectory, growth rate, ...)
    - Based on statistical inference and (structural) models
    - Model specification uses additional information from an underlying research field

Model for numerical versus categorical data

- Choice of a method for analyzing longitudinal data depends on
  - The research objective, Do you want to create a typology, to discover association among your variables, or do you want to test some effects (gender, birth cohorts, education,...)?
  - Are you interested in trajectories or in specific transitions?
  - The kind of longitudinal data you have: Questions and methods for numeric variables are quite different from those for categorical variable.

What can we do with longitudinal data?

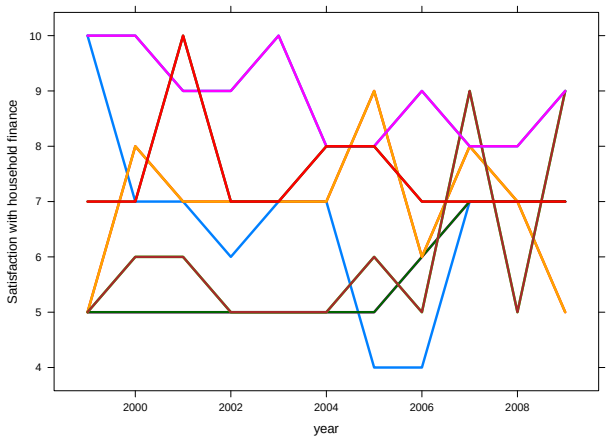
- Longitudinal data track individuals over time.
- By accounting for the individual time dimension, they allow for analysis of individual dynamics (changes over time)
  - life trajectories (cohabitational, occupational, health states, ...)
  - transitions, social mobility, ...
  - growth of ability indexes, financial resources,
- By considering multiple units, they also permit to analyse differences, or heterogeneity, among individual life courses.

Numerical Longitudinal Data

- Exploratory methods:
  - Sequence of transversal characteristics: e.g.; means, medians, between subject standard deviations, ...
  - Summary of longitudinal characteristics: e.g.; mean, median, standard deviation, within subject standard deviations, ...
  - Clustering individual series: e.g.; k-means for longitudinal data, (Genolini and Falissard, 2010)
  - Regression trees for longitudinal data: e.g.; RE-EM Trees, (Sela and Simonoff, 2009)
  - Graphical display: e.g.; spaghetti plots, ...

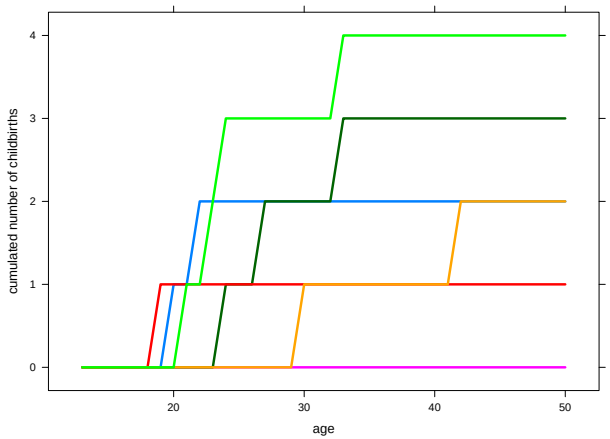
Spaghetti plot: Satisfaction with HH income

10 households



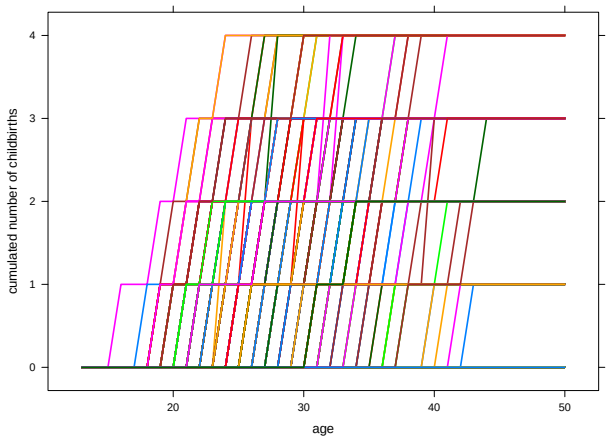
Spaghetti plot: Childbirth trajectories

Fertility trajectories of 6 randomly selected women



Spaghetti plot: Childbirth trajectories

Fertility trajectories of a sample of 500 women of the 1945-49 birth cohort



Regression models for numerical longitudinal data

- Advanced modeling techniques for studying how the individual dynamics (trajectories, growth rate, changes, ...) are related to the (historical and socio-economic) context as well as to individual resources.
- Regression-like models linking the trajectories to covariates.

$$y_{it} = \beta_{0,i} + \beta_{1,i}x_{1,it} + \cdots + \beta_{2,i}x_{2,it} + \epsilon_{it}$$

Evolution of Divorce and AFDC, US states

Illustration of random effect

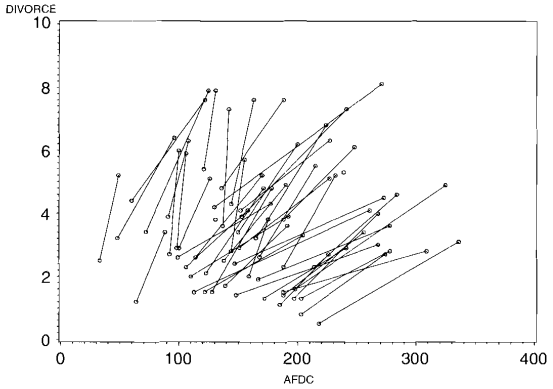


Figure 1.2. Plot of divorce rate versus AFDC payments from 1965 and 1975.

Plot from Frees (2004, p 4), Units are 50 US states,

AFDC is Aid to Families with Dependent Children

Longitudinal models

- Aim: Measure the common underlying growth by accounting for the heterogeneity among subjects
  - Fixed effect model: same $\beta_{k,i}$ for all individuals i
    - Accounts for heterogeneity with a dummy for each individual
    - only works with a small number of units
  - Random effect model: random coefficient: $\beta_{ik} = \beta_k + \alpha_i$, with α_i a random term (shared by all observations over time of a same unit)
  - Mixed effect model: Fixed and random effects
- The above models can be seen as special cases of multilevel models.

Alternative simple linear models

in presence of G units i

Model		Stdev of u	# coef.	
$m1$	$y_{it} = a + bx_{it} + u_{it}$	σ	$2+1$	average model
$m2$	$y_{it} = a_i + b_i x_{it} + u_{it}$	$\sigma_1 \ldots \sigma_i$	$G(2+1)$	independent
$m3$	$y_{it} = a_i + b_i x_{it} + u_{it}$	σ	$2G+1$	seemingly indep.
$m4$	$y_{it} = a_i + bx_{it} + u_{it}$	σ	$G+2$	dummies
$m5$	$y_{it} = (a + u_{ai}) + (b + u_{bi})x_{it} + u_{it}$	$\sigma_a, \sigma_b, \sigma$	$2+3$	random effects
$m6$	$y_{it} = (a + u_{ai}) + bx_{it} + u_{it}$	σ_a, σ	$2+2$	shared frailty

Models with latent variables

- Models to study evolution of an non-observed latent variable
  - Latent curve model
  - Latent change score model
  - Multi-occasion latent trait models (categorical manifest variables)
- More details in Paolo Ghisletta's presentation)

Categorical responses

- Categorical responses are either states or events
  - states last
  - events just occur at a given time point
- State sequences: series of cohabitational, occupational, health, ... states
  - (2P, 2P, 2P, A, A, AC, AC, MC, MC, ...)
- Event sequences (event histories): successive time stamped events
  - (Left home at 20, Childbirth at 22, Married at 24, ...)

Holistic versus Event specific approaches

- Holistic approach: the sequence is considered as a whole unit
  - State sequence analysis: What are typical life course patterns?
  - Event sequence analysis: What is the typical sequencing of life events?
  - Markov models: How does current state (current event occurrence) depend on previous states (event occurrences)?
- Event specific approach: focus is on a given event, or transition
  - Survival analysis (also known as Event history analysis):
    - Analysis of the duration (time to event);
    - Hazard of experiencing the event after a time t
    - How is the duration (or hazard) linked to covariates?

Longitudinal categorial response: causal vs exploratory

- Sequence analysis is mainly exploratory
- Survival analysis is more causal oriented
- Modeling, especially latent class (categorical latent variable) or latent trait (categorical manifest variables) models, can also be extended for longitudinal variables
- More details in presentation on Categorical response

Insight to IP 14
Measuring life sequences and the disorder of lives

- Aim of IP 14
  - provide LIVES with high methodological competencies in advanced statistical tools used in the longitudinal analysis of life courses
  - and develop methods specifically suited for describing and investigating vulnerability processes.
- IP 14 is not a help desk!
  - IP14 will collaborate with other IP's to find the best suited solutions for their research objectives
  - It will not run analyses for the IP's, however!

Insight to IP 14
Vulnerability indicators

Vulnerability indicators

- Identify sequence structures that reflect vulnerability (build on sequence complexity indicators, Gabadinho et al. 2010)
- Vulnerability as latent variable revealed by atypical behaviors (suitable for analyzing the evolution of vulnerability variance along the life course)
- Hidden Markov Models for analyzing how a switch in a hidden vulnerability state can increase probability to fall in unwanted situations such as unemployment or broken social network.

Insight to IP 14
Vulnerability trajectories

Vulnerability trajectories

- Identify vulnerability sequences as atypical sequences that depart (in some way) from standard ones.
- Measure degree of vulnerability through distance to closest 'regular' representative trajectory
- Identify unusual order of life events that may increase later probability to be confronted to disruptive events
- Confront parallel trajectories (familial and professional, linked lives, ...) should also permit to highlight relationship between trajectory and vulnerability

Insight to IP 14
Vulnerability and its context

Vulnerability and its context

Next step will be to study how vulnerability is linked to contextual variables variables that characterize the social and economic environment as well personal resources.

- Regression models using vulnerability indicators
- SEM models for relation of latent vulnerability and context
- Survival methods for studying hazard to become vulnerable

Insight to IP 14

Disadvantages accumulation and stress proliferation

Disadvantages accumulation and stress proliferation

- How can disadvantage accumulation (Dannefer, 2003) be empirically measured and tested?

Thank You!

References I

Beck, N. and J. N. Katz (1995). What to do (and not to do) with time-series cross-section data. *American Political Science Review 89*, 634–647.

Dannefer, D. (2003). Cumulative advantage, and the life course: Cross-fertilizing age and social science knowledge. *Journal of Gerontology 58b*(6), S327–S337.

Elder, G. H. (1974). *Children of the great depression. Social change in life experience. Chicago: Univ. of Chicago Press*. Chicago: Univ. of Chicago Press.

Frees, E. W. (2004). *Longitudinal and Panel Data: Analysis and Applications in the Social Sciences*. Cambridge.

Gabadinho, A., G. Ritschard, N. S. Müller, and M. Studer (2011). Analyzing and visualizing state sequences in R with TraMineR. *Journal of Statistical Software 40*(4), 1–37.

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2010). Indice de complexité pour le tri et la comparaison de séquences catégorielles. *Revue des nouvelles technologies de l'information RNTI E-19*, 61–66.

References II

Gabadinho, A., G. Ritschard, M. Studer, et N. S. Müller (2011). Extracting and rendering representative sequences. In A. Fred, J. L. G. Dietz, K. Liu, et J. Filipe (Eds.), *Knowledge Discovery, Knowledge Engineering and Knowledge Management*, Volume 128 of *Communications in Computer and Information Science (CCIS)*, pp. 94–106. Springer-Verlag.

Genolini, C. et B. Falissard (2010). KmL : k-means for longitudinal data. *Computational Statistics 25*(2), 317–328.

Giele, J. Z. et G. H. Elder, Jr. (1998). *Methods of Life Course Research : Qualitative and Quantitative Approaches*. Thousand Oaks, CA : Sage.

Perroux, O. et M. Oris (2005). Présentation de la base de données de la population de Genève de 1816 à 1843. Séminaire statistique sciences sociales, Université de Genève. (http ://mephisto.unige.ch/pub/stats/seminaire/early_life.pdf).

Ritschard, G., A. Gabadinho, N. S. Müller, and M. Studer (2008). Mining event histories: A social science perspective. *International Journal of Data Mining, Modelling and Management 1*(1), 68–90.

References III

Ritschard, G., A. Gabadinho, M. Studer, and N. S. Müller (2009). Converting between various sequence representations. In Z. Ras and A. Dardzinska (Eds.), *Advances in Data Management*, Volume 223 of *Studies in Computational Intelligence*, pp. 155–175. Berlin: Springer-Verlag.

Sela, R. J. and J. S. Simonoff (2009). RE-EM trees: A new data mining approach for longitudinal data. IOMS Statistics Working Papers SOR-2009-03, New York University.

Singer, J. D. and J. B. Willett (2003). *Applied longitudinal data analysis: Modeling change and event occurrence*. Oxford: Oxford University Press.

Wanner, P. et E. Delaporte (2001). Reconstitution de trajectoires de vie à partir des données de l'état civil (BEVNAT). une étude de faisabilité. Rapport de recherche, Forum Suisse des Migrations.

Wernli, B. (2010). A Swiss survey landscape for communication research. In *Università della Svizzera Italiana, USI, Lugano, 2010, June 15, Institute of Communication and Health*.