Some inference tricks in event history analysis

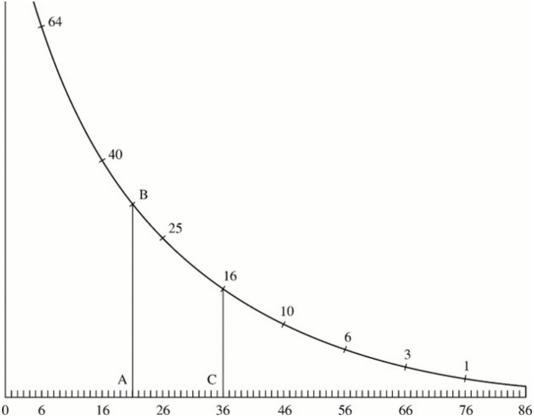
Ernst C. Wit

Università della Svizzera italiana,
Lugano, Switzerland

Joint work with Ruta Juozaitiene, Martina Boschi, Edoardo Filippi-Mazzola

1 February 2024

Christiaan Huygens early survival analysis



Huygens' 1669 survival curve for 100 people.

Howard Wainer STATISTICAL GRAPHICS: Mapping the Pathways of Science.

Annual Review of Psychology. Vol. 52: 305-335.

Università della Svizzera italiana

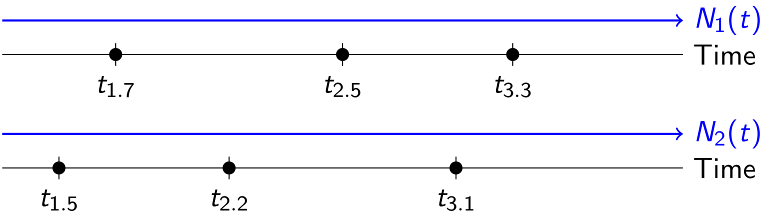
Event history analysis 101

Università
della
Svizzera
italiana

Counting Process Formulation of Event History Process

Event history process models occurrence of events over time.

- Event Time t: observation time.
- Event Indicator $\delta_i(t)$: indicator whether event i occurred at time t.



Mathematical Formulation:

Given observations from countable marked point process $\{T_j\}_{j=1}^n$,

$$N_i(t) = \sum_{j=1}^{n} \delta_i(t_j)$$

$N_i(t) =$ cumulative count of events of type i up to time t.

Università della Svizzera italiana

Doob-Meyer Decomposition

Doob-Meyer decomposes a general stochastic process into

- a predictable process (the "model").
- and a martingale ("noise")

Doob-Meyer Decomposition of counting process N_i:

$$N_i(t) = M_i(t) + \Lambda_i(t)$$

- $M_i(t)$: Martingale component representing the unpredictable part.
- $\Lambda_i(t)$: Predictable component capturing the systematic part.

Common assumption:
Events cannot happen simultaneously a.s., and therefore

$$\Lambda_i(t) = \int_0^t \lambda_i(s)\, ds$$

cumulative hazard written as a integral of some hazard function.

Università della Svizzera italiana

Cox Proportional Hazards Model

- Introduced by David R. Cox in 1972.
- Allows for analysis of effect of covariates on hazard function.

Mathematical Formulation:

$$\lambda(t|x) = Y_i(t)\lambda_0(t)e^{\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k}$$

where:

- $\lambda(t|x)$: Hazard function at time t given covariates x.
- $Y_i(t)$: Indicator whether event type i is at risk at time t.
- $\lambda_0(t)$: Baseline hazard function.
- $\beta_1, \beta_2, \ldots, \beta_k$: Regression coefficients.
- $x_1, x_2, \ldots, x_k$: Covariates.

Università della Svizzera italiana

Partial Likelihood Formulation

Full likelihood of observed event types $\{i_j\}_{j=1}^n$ at times $\{t_j\}_{j=1}^n$,

$$L = \prod_{j=1}^{n} p(i_j, t_j \mid i_{<j}, t_{<j})$$

Università della Svizzera italiana

Partial Likelihood Formulation

Full likelihood of observed event types $\{i_j\}_{j=1}^n$ at times $\{t_j\}_{j=1}^n$,

$$
\begin{aligned}
L &= \prod_{j=1}^{n} p(i_j, t_j \mid i_{<j}, t_{<j}) \\
&= \prod_{j=1}^{n} p(t_j \mid i_{<j}, t_{<j}) \times P(i_j \mid i_{<j}, t_{\leq j})
\end{aligned}
$$

Università della Svizzera italiana

Partial Likelihood Formulation

Full likelihood of observed event types $\{i_j\}_{j=1}^n$ at times $\{t_j\}_{j=1}^n$,

$$
\begin{aligned}
L &= \prod_{j=1}^{n} p(i_j, t_j \mid i_{<j}, t_{<j}) \\
&= \prod_{j=1}^{n} p(t_j \mid i_{<j}, t_{<j}) \times P(i_j \mid i_{<j}, t_{\leq j}) \\
&= \prod_{j=1}^{n} P(t_j \mid i_{<j}, t_{<j}) \times \frac{\lambda_{i_j}(t_j)}{\sum_{i \in \mathcal{R}(t_j)} \lambda_i(t_j)}
\end{aligned}
$$

Università della Svizzera italiana

Partial Likelihood Formulation

Full likelihood of observed event types $\{i_j\}_{j=1}^n$ at times $\{t_j\}_{j=1}^n$,

$$\begin{aligned} L &= \prod_{j=1}^{n} p(i_j, t_j \mid i_{<j}, t_{<j}) \\ &= \prod_{j=1}^{n} p(t_j \mid i_{<j}, t_{<j}) \times P(i_j \mid i_{<j}, t_{\leq j}) \\ &= \prod_{j=1}^{n} P(t_j \mid i_{<j}, t_{<j}) \times \frac{\lambda_{i_j}(t_j)}{\sum_{i \in \mathcal{R}(t_j)} \lambda_i(t_j)} \end{aligned}$$

The **Partial Likelihood** for Cox PH is defined as:

$$L_P(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\sum_{i \in \mathcal{R}(t_j)} \exp(\beta^T x_i)}$$

where $\mathcal{R}(t_j) = \sum_{i=1}^{p} Y_i(t_j)$ is risk set at time t_j.

Challenges and Tricks

Università
della
Svizzera
italiana

Challenges in Inference of Event History Models

- **Growing Risk Set:** for big studies with many event types.
- **Time-varying covariates:** requires artificial censoring in software.

ELSEVIER

Child Abuse & Neglect
Volume 28, Issue 9, September 2004, Pages 947-966

An event history analysis of recurrent child maltreatment reports in Florida

Lodi Lipien [a], Melinda S. Forthofer [b]

Show more

+ Add to Mendeley Share Cite

https://doi.org/10.1016/j.chiabu.2004.03.011 Get rights and content

Abstract

Objective:

The purpose of this study was to (a) describe the timing of maltreatment recurrence and (b) measure associations between child demographics and characteristics of initial reports with recurrent maltreatment.

Method:

Using administrative data from the Florida Department of Children and Families, case histories of 189,375 children with an initial maltreatment report in 1998 or 1999 were

Source on Save Run

```r
# Introduce censoring for time-dependent covariate
example <- list(start = c(1, 2, 4),
                stop  = c(2, 4, 8),
                event = c(0, 0, 1),
                x     = c(1, 0, 2))
```

- **Non-linear effects:** either $x_i\beta(t)$ or $f(x_i)$
- **Global covariates:** either $x\beta$ or λ_0

Università della Svizzera italiana

TRICK 1: nested-case control sampling

When $\mathcal{R}(t)$ gets too big, then Borgan (1995) suggests alternative:

$$L_{NCC}(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\sum_{i \in \mathcal{S}(t_j)} \exp(\beta^T x_i)}$$

where $\mathcal{S}(t_j)$ is a randomly sampled subset of $\mathcal{R}(t_j)$ *including* event i_j.

Università della Svizzera italiana

TRICK 1: nested-case control sampling

When $\mathcal{R}(t)$ gets too big, then Borgan (1995) suggests alternative:

$$L_{NCC}(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\sum_{i \in \mathcal{S}(t_j)} \exp(\beta^T x_i)}$$

where $\mathcal{S}(t_j)$ is a randomly sampled subset of $\mathcal{R}(t_j)$ *including* event i_j.

If we sample **only 1 non-event** i_j^* for each event i_j, then

$$L_{NCC}(\beta) \quad = \quad \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\exp(\beta^T x_{i_j}) + \exp(\beta^T x_{i_j^*})}$$

TRICK 1: nested-case control sampling

When $\mathcal{R}(t)$ gets too big, then Borgan (1995) suggests alternative:

$$L_{NCC}(\beta) = \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\sum_{i \in \mathcal{S}(t_j)} \exp(\beta^T x_i)}$$

where $\mathcal{S}(t_j)$ is a randomly sampled subset of $\mathcal{R}(t_j)$ *including* event i_j.

If we sample **only 1 non-event** i_j^* for each event i_j, then

$$\begin{aligned} L_{NCC}(\beta) &= \prod_{j=1}^{n} \frac{\exp(\beta^T x_{i_j})}{\exp(\beta^T x_{i_j}) + \exp(\beta^T x_{i_j{}^*})} \\ &= \prod_{j=1}^{n} \frac{\exp(\beta^T (x_{i_j} - x_{i_j{}^*}))}{1 + \exp(\beta^T (x_{i_j} - x_{i_j{}^*}))} \end{aligned}$$

Logistic regression

This is logistic regression with only successes and covariates $\Delta x_i = x_i - x_{i^*}$.

TRICK 2: Generalized additive model for event history data

Consider we want to model hazard as

$$\lambda_i(t) = \lambda_0(t) e^{x_{i1}(t)\beta(t) + f(x_{i2}(t)) + z_i(t)\gamma}$$

with

- time-varying covariates $x_{i1}(t), x_{i2}(t), z_i(t)$.
- time-varying effect $\beta(t)$
- non-linear effect f
- random effect $\gamma \sim N(0, \sigma^2)$ (frailty)

with observed data $\{(i_j, t_j)\}$ with $j = 1, \ldots, n$.

TRICK 2: practical approach

1. For each event i_j, sample one non-event i_j^* from $\mathcal{R}(t_j)$.
2. Create new covariates:
   - $\Delta z_j = z_{i_j} - z_{i_j^*}$
   - $\Delta x_{1j} = x_{i_j 1} - x_{i_j^* 1}$
   - matrices $\Delta x_2 = \begin{bmatrix} x_{i_1 2} & x_{i_1^* 2} \\ \vdots & \vdots \\ x_{i_n 2} & x_{i_n^* 2} \end{bmatrix}$ and $C = \begin{bmatrix} 1 & -1 \\ \vdots & \vdots \\ 1 & -1 \end{bmatrix}$
3. time variable $t = (t_1, \ldots, t_n)$
4. pseudo response $y = c(1, \ldots, 1)$

In R using mgcv package, we can now fit the event history model:

```
library(mgcv)
model <- gam(y ~ s(t,by=DeltaX1) + s(DeltaX2, by=C)+ s(DeltaZ,bs="re"),
             family = binomial)
```

Application: Alien Species Invasions

Università della Svizzera italiana

Motivation: species invasions

Alien species are increasingly recognized as threat to native ecology.

Question

What are the main drivers of the invasive process of species?

Data: Alien Species First Records

Primary data source: Alien Species First Records (Seebens et al., 2018)

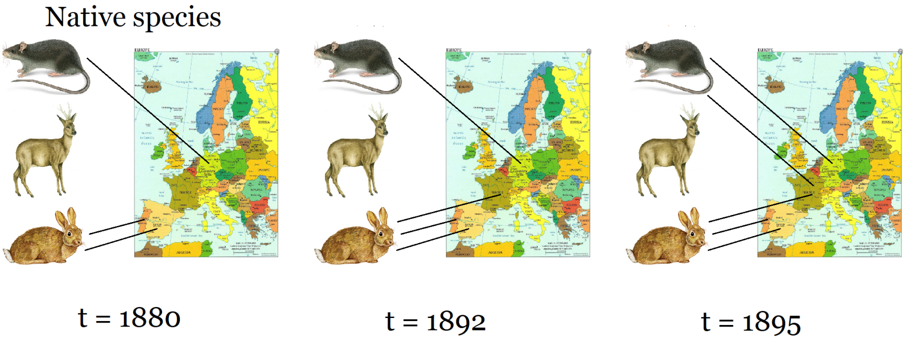
H	I	J	K	L	M	N	O	P
OrigName	LifeForm	Region	PresentStatus	FirstRecor	FirstRecor	DataQualit	Source	
Acanthophora muscoides Linnaeus, 175	Algae	Turkey		1986	1986		Cinar et al. (2005)	
Acanthophora nayadiformis	Algae	Cyprus	alien	1997	1997	NEW_Befo	DAISIE	
Acanthophora nayadiformis (Delile) Pap	Algae	Turkey		1970	1970		Cinar et al. (2005)	
Acanthophora spicifera	Algae	Hawaiian Islands		1952	1952		Carlton & Eldrege (2009)	
Acetabularia calyculus	Algae	Israel	established	1943	1943	NEW_Befo	DAISIE	
Acetabularia calyculus	Algae	Spain	established	1957	1957	NEW_Befo	DAISIE	
Achnanthes pseudogroenlandica	Algae	Bulgaria		1984	1984		aquaNIS	
Achnanthes pseudogroenlandica	Algae	Romania		1984	1984		aquaNIS	
Achnanthes pseudogroenlandica	Algae	Ukraine		1984	1984	NEW_Befo	DAISIE	
Acrochaetium catenulatum	Algae	Netherlands		1967	1967		aquaNIS	
Acrochaetium kylinii	Algae	Turkey		2007	2000 - 2009	NEW_rand	aquaNIS	
Acrochaetium leptonema	Algae	Bulgaria		2006	2000 - 2009	NEW_rand	aquaNIS	
Acrochaetium leptonema	Algae	Turkey		2007	2000 - 2009	NEW_rand	aquaNIS	

Effectively giving information for

1. for each **species** (inside 1 of 16 life forms)
2. for each **"region"** (of 275 regions)
3. **First moment** that species is recorded there.

Università della Svizzera italiana

Dynamic two-mode species-region network



In a dynamic two-mode species-region network:

- **species** and **regions** are node-sets;
- Edges in network are **time-stamped invasions**;
- At time 0 ($t = 1880$) native species are indicated by edge;
- Invasions **may** show spatial trend or co-occuring patterns.

Università della Svizzera italiana

Event history model for time-to-invasion

Formally, **data** for all species $s \in \mathcal{S}$ and regions $r \in \mathcal{C}$:

$$T_{sr} = \text{year in which species } s \text{ appeared in region } r.$$

- native species: $T_{sr} < 1880$
- invasions: $T_{sr} \in [1880, 2005]$
- non-invasions: $T > 2005$

Define a **Cox proportional hazards model**:

$$\lambda_{sr}(t) = \text{hazard of species } s \text{ invading region } r \text{ in year } t.$$

by means of

$$\lambda_{sr}(t) = \lambda_0(t) e^{x'_{sr}(t)\beta(t) + z'_{sr}(t)\gamma}$$

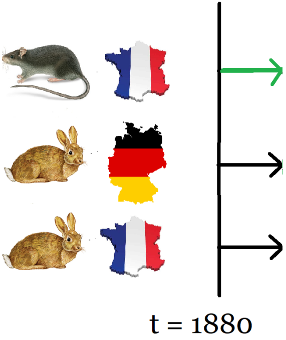
where $\lambda_0(t)$ baseline hazard, $x_{sr}(t)$ fixed effect, $z_{sr}(t)$ random effect.

Università della Svizzera italiana

Idea behind event history model

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all species for life form $L =$ mammals.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

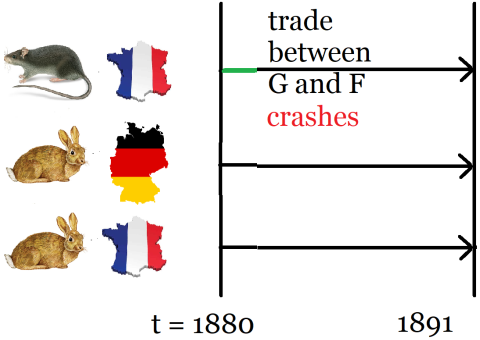
We consider a "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Idea behind event history model

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all species for life form $L = \text{mammals}$.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

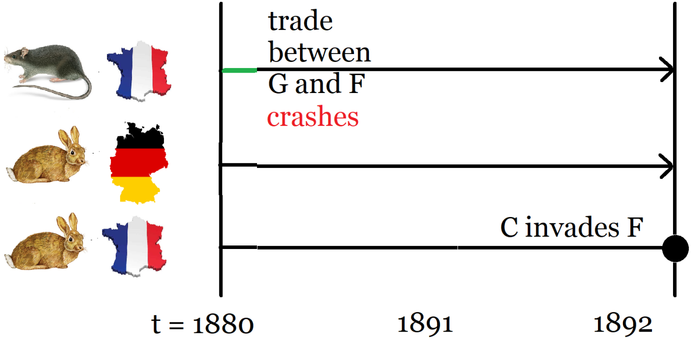
We consider a “race” between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Idea behind event history model

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all species for life form $L = \text{mammals}$.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

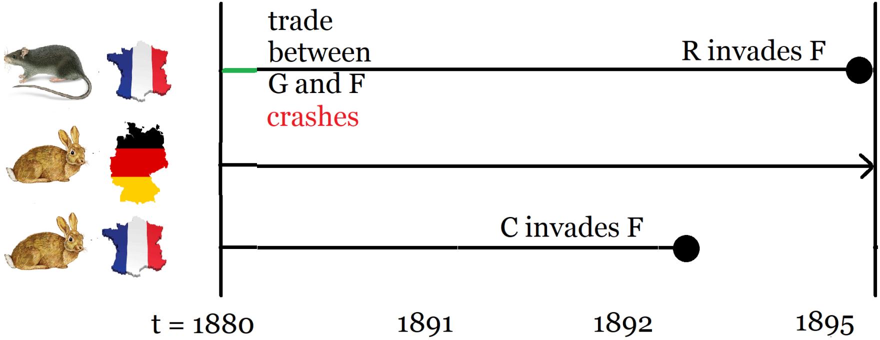
We consider a "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Idea behind event history model

- Let $\mathcal{S}_L = \{\text{rattus, cuniculus}\}$ be all species for life form $L = \text{mammals}$.
- Let $\mathcal{C} = \{\text{Germany, France}\}$ be all regions.
- Let *rattus* be native to *Germany*.

We consider a "race" between T_{rF}, T_{cG} and T_{cF} (T_{rG} already arrived!):



Possible drivers of species invasions: fixed effects

Most drivers change in time:

- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: diff in temperature between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.

$d_{rF}(1880) = 780\text{km}$

Possible drivers of species invasions: fixed effects

Most drivers change in time:

- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: diff in temperature between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.

$d_{rF}(1880) = 780\text{km}$

$d_{rF}(1883) = 780\text{km}$

Università della Svizzera italiana

Possible drivers of species invasions: fixed effects

Most drivers change in time:

- $l_r(t)$: landuse in region r at time t.
- $d_{sr}(t)$: distance to region nearest to r invaded by s by time t.
- $tr_{sr}(t)$: annual trade between r and regions invaded by s by time t.
- $dt_{sr}(t)$: diff in temperature between r and regions invaded by s by time t.
- $k_{sr}(t)$: presence of s at time t in colonial power to which r belongs.

$d_{rF}(1880) = 780$km $d_{rF}(1883) = 780$km $d_{rF}(1889) = 570$km

Università della Svizzera italiana

Possible drivers of species invasions: random effects (I)

Random effects are used for large number of “nuisance” factors:

- **Invasiveness.** Different species may vary in their invasive behaviour, *beyond* fixed effects. We model **species invasiveness** by:

$$\gamma_s \sim N(0, \sigma^2_{\text{inv}}).$$

- **Popularity.** Certain regions may be more “popular” destinations than others, *beyond* fixed effects. We model **region popularity** by:

$$\gamma_c \sim N(0, \sigma^2_{\text{pop}}).$$




Possible drivers of species invasions: random effects (II)

Species interaction network

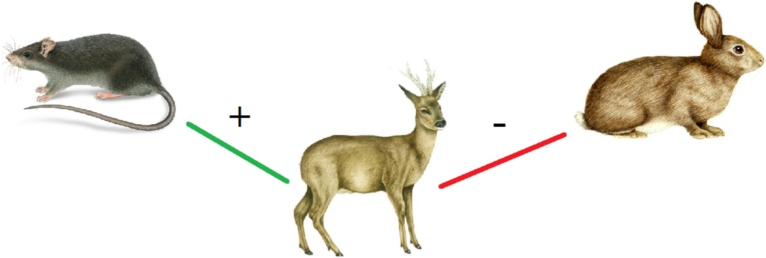
Could it be that certain species co-invade a region?
Or, reversely, avoid each other in their invasions?

We define:

$$i_c(t) = \text{last species to invade } r \text{ before } t.$$

and

$$\gamma_{ss'} = \text{affinity of species } s \text{ for species } s'$$



Estimation

There are a number of estimation paradigms:

- **MLE:** Computationally intractable even for small networks.
- **Partial likelihood:** Denominator of PL is sum of $|\mathcal{S}| \times |\mathcal{C}|$ terms.

Case-control Partial Likelihood:

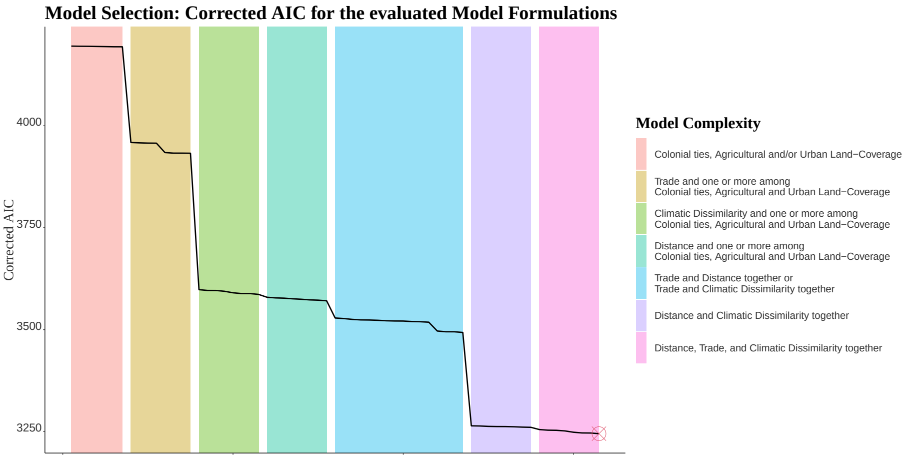
Randomly sample 1 non-event (t_i, s_i^*, r_i^*) for each event (t_i, s_i, r_i).

$$(\hat{\beta}, \widehat{\Sigma}_\gamma) = \text{argmin} \prod_{i=1}^{n} \frac{e^{x_{s_i r_i}\beta + z_{s_i r_i}\gamma}}{e^{x_{s_i r_i}\beta + z_{s_i r_i}\gamma} + e^{x_{s_i^* r_i^*}\beta + z_{s_i^* r_i^*}\gamma}}$$

This is equivalent with **logistics regression**

- for which responses are ones, $y = (1, 1, \ldots, 1)$
- for which covariates are differences, $x_{s_i r_i} - x_{s_i^* r_i^*}$ and $z_{s_i r_i} - z_{s_i^* r_i^*}$.

Model selection



Università della Svizzera italiana

Results: fixed effects

	Birds	Plants	Insects	Mammals
Colonial ties	0.16	-0.09	0.31	0.13
Difference in temperature	-0.08	-0.04	-0.11	-0.07

Climatic effect

All life forms diffuse to "similar climatic" regions:

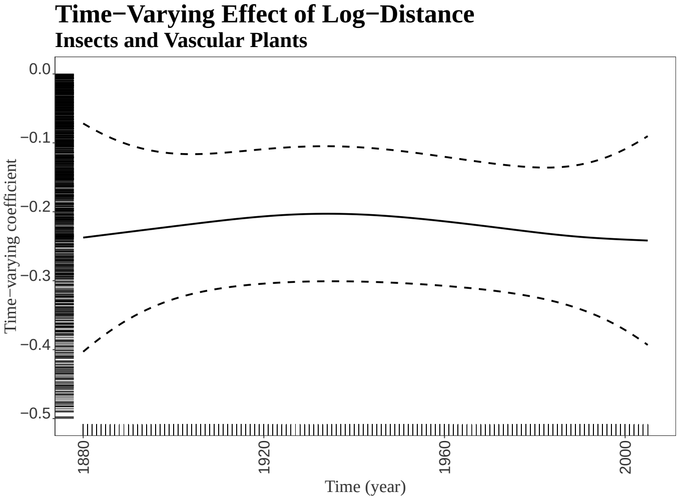
1. Strongest for *insects*
2. Weakest for *plants*

Colonial history

Insects: **less** diffusion among countries related by colonial history.
Birds, insects and mammals: there is **more**.

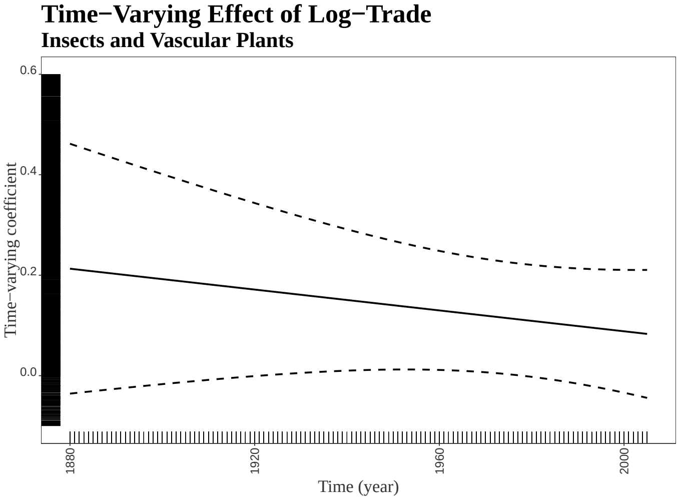
Università della Svizzera italiana

Results: distance reduces invasions



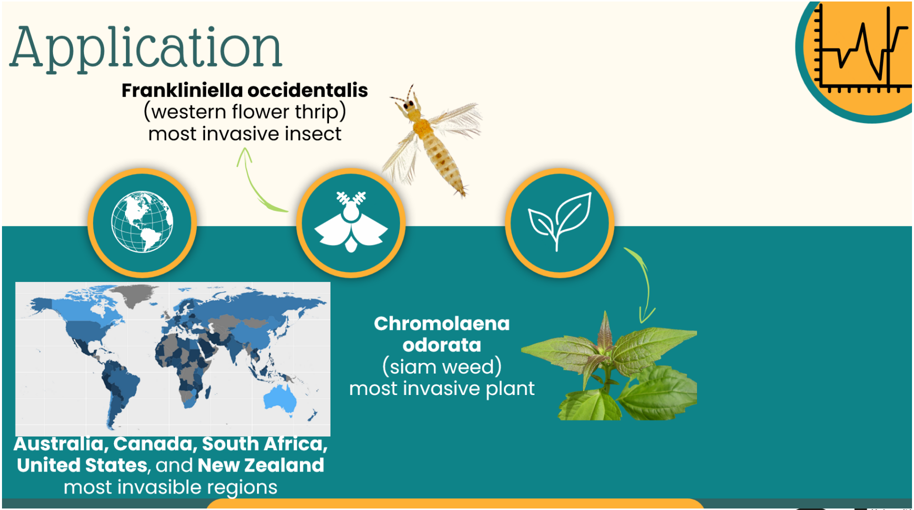
Distance effect is negative and quite constant over period 1880-2005.

Results: trade is becoming less important



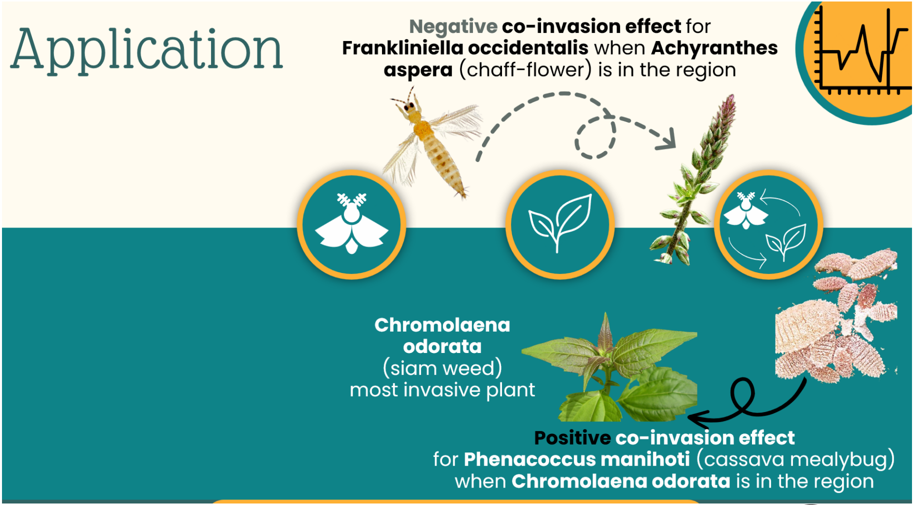
Trade effect is positive, but decreasing over period 1880-2005.

Results: Random effects for species and countries



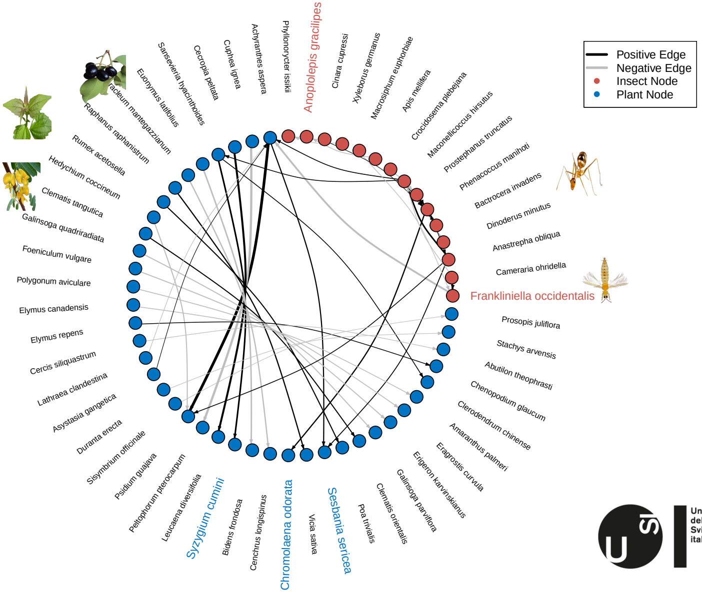
Università
della
Svizzera
italiana

Results: Random interaction effects between species



Università della Svizzera italiana

Results: Species have a tendency to coinvade



Summary

- **Event history analyis** is an important tool in Biostatistics
- **Computational tricks** model big data in a more realistic way:
  - **Nested-case control** improves computational efficiency
  - **Logistic formulation** allows use of non-linear modelling via GAMs.
- Species invasions as **temporal two-mode dynamic process**.
  - Surprising similarities in dynamics for various life forms.
  - Trade, geographic and climatic distance are important drivers.
  - Significant variation in invasiveness of species and regions.
  - Species networks hint at joint invasion dynamics.