Analyse de trajectoires par analyse des correspondances fonctionnelle

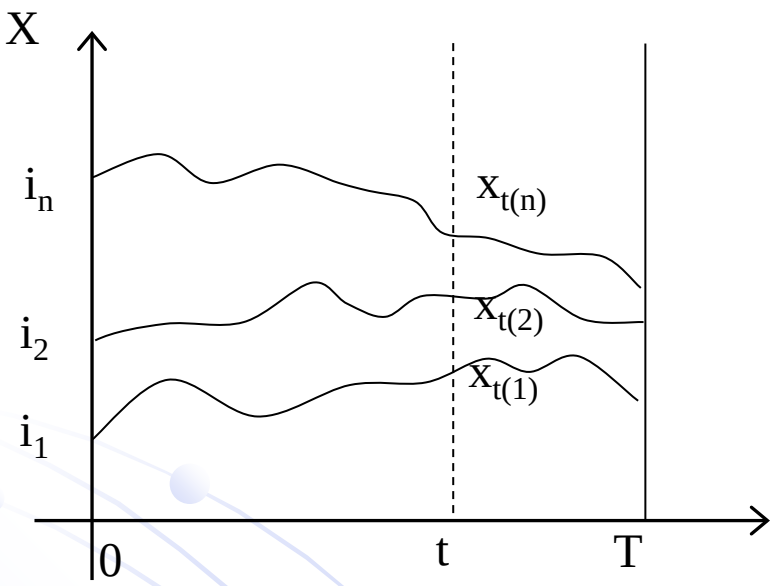
Gilbert Saporta
Chaire de Statistique Appliquée
Conservatoire National des Arts et Métiers
292 Rue Saint Martin
75141 Paris Cedex 03
gilbert.saporta@cnam.fr

INTRODUCTION

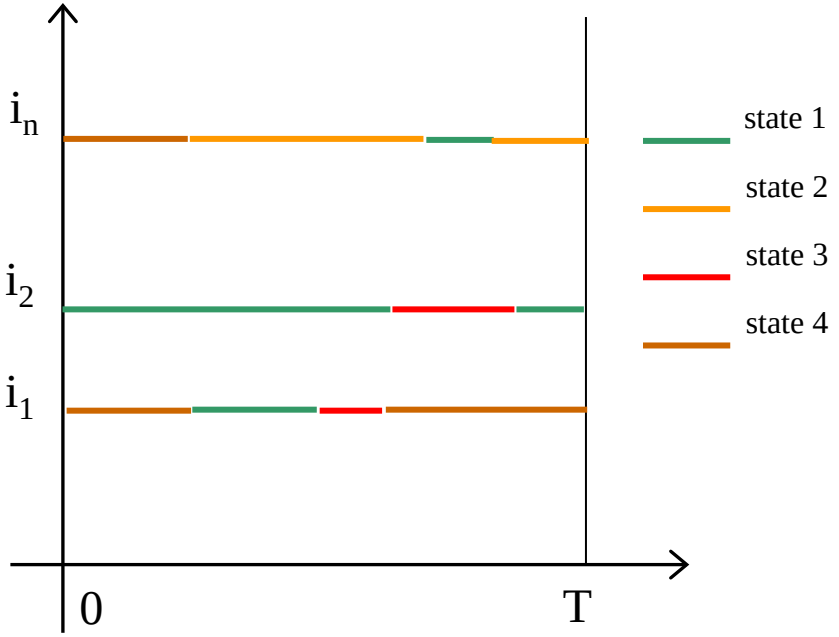
- Premiers travaux:
  - J. C. Deville – 1974
  - P. Besse – 1979
  - G. Saporta – 1981
- Ensuite...
  - Aguilera, Valderrama – 1993, 1995, 1998
  - Ramsay, Silverman – 1995, 1997
  - Van der Heijden – 1997
  - Preda, Cohen – 1999
  - Cardot, Ferraty, Vieu - 1999, 2005

Données fonctionnelles

NUMERIQUES



QUALITATIVES



« event history data »

PROCESSUS NUMERIQUE

Exemples:

- Taille d'une famille t années après le mariage
- Valeur boursiere

Pour chaque t

variable numérique:

$$x_t = \begin{pmatrix} x_t(1) \\ . \\ . \\ . \\ x_t(n) \end{pmatrix}$$

Infinité non dénombrable de variables si $t \in [0;T]$

PROCESSUS QUALITATIF

Exemples:

- Phases du sommeil
- Statut social
- Statut matrimonial

A chaque instant t
variable nominale x_t à m catégories.

I ACP FONCTIONNELLE

X_t centré $\quad X_t \in L^2(\Omega \times T)$

- fonction de covariance: $C(t,s) = E(X_t X_s)$
- Opérateur de covariance C

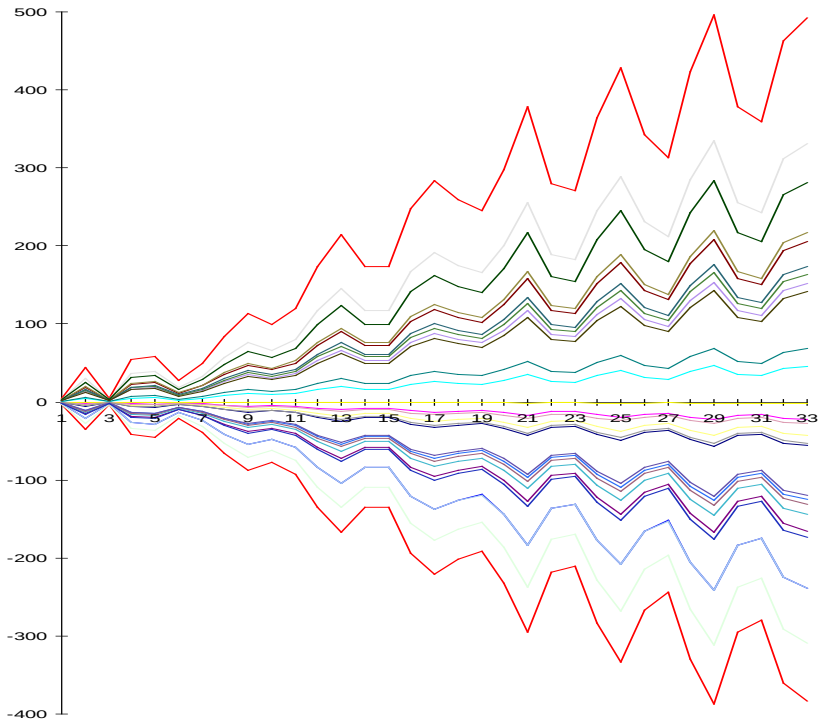
$$f(s) \to \int_0^T C(t,s) f(s) ds$$

- Combinaison linéaire $\quad \xi = \int_0^T f(t) X_t dt$

I.1 Processus quasi – deterministe

$$X_t(i) = \xi_i f(t)$$

Même forme à une constante près ξ_i relative à l'individu i



Tout processus peut être approché par une somme de processus quasi det.

$$X_t \# \sum_k \xi^k f^k(t)$$

$\xi_i^{(k)}$=coordonnée sur l'axe k

Choix de la base $f^k(t)$

Fonctions orthogonales de $L^2(T)$

$$\int_0^T f^k(t)\, f^l(t) dt = \begin{cases} 1 if\ k = l \\ 0 if\ k \neq l \end{cases}$$

Fourier par exemple:

$$f^k(t) = \cos\frac{2k\pi t}{T} \ or \ \sin\frac{2k\pi t}{T} \Rightarrow \xi_k = \int_0^T X_t\, f^k(t) dt$$

MAIS les ξ_k sont corrélés.

I.2 Décomposition de Karhunen – Loeve

Décomposition unique

$$X_t = \sum_{k=1}^{\infty} \xi_k f_k(t)$$

f_k= ensemble orthonormé de fonctions de $L^2(T)$

ξ_k = ensemble orthogonal (non-corrélation) de variables de $L^2(\Omega)$

f_k fonctions propres de C

$$\lambda_k f_k(t) = \int_0^T C(t,s) f_k(s) ds \qquad V(\xi_k) = \lambda_k$$

ξ_k fonctions propres de W

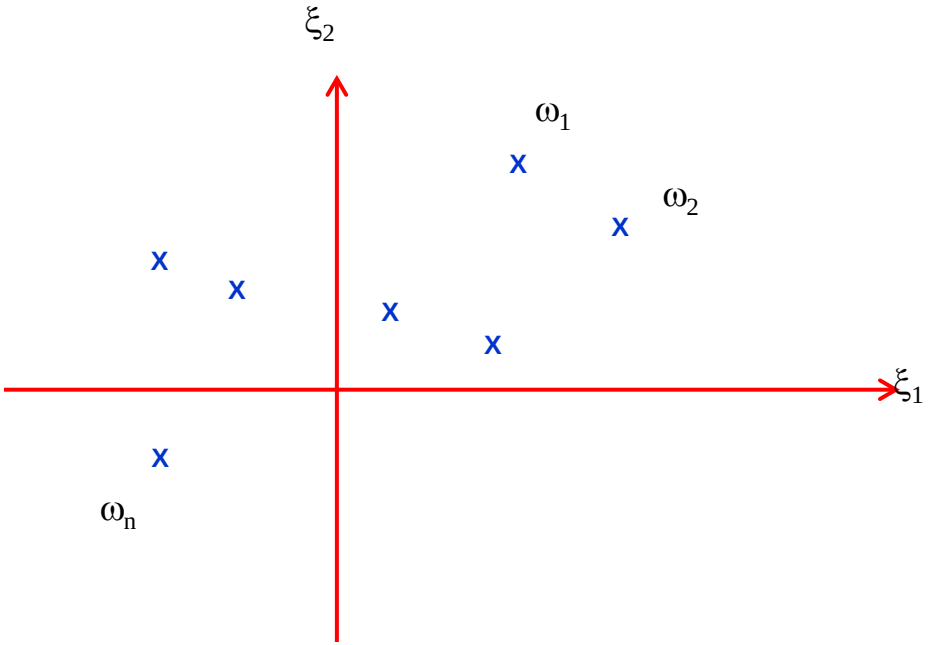
$$\xi_k = \int_0^T X_t f_k(t) dt$$

I.3 décomposition de Karhunen–Loeve et ACP

f_k facteurs principaux

ξ_k comosantes principales (coordonnées sur l'axe k)

KARHUNEN – LOEVE ≡ SVD (SINGULAR VALUE DECOMPOSITION)



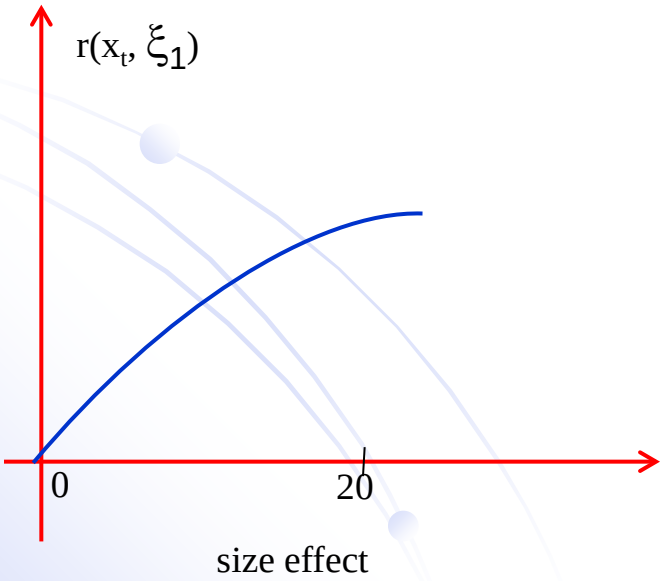
Coordonnées indépendantes de t; facteurs fonctions de t

Interprétation:

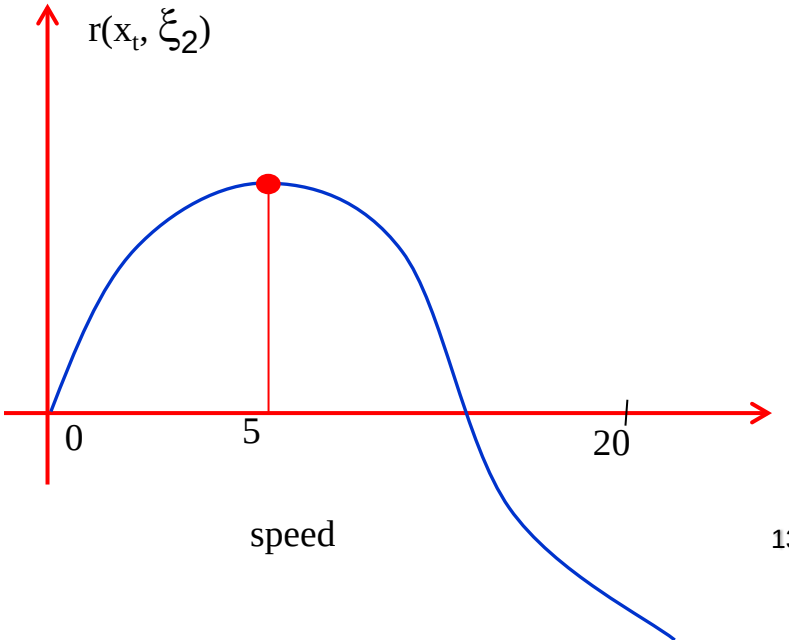
Comme en ACP

$$r\left(X_t;\xi_k\right)=\frac{\sqrt{\lambda_k}\,f_k(t)}{\sigma(t)}$$

Exemple: taille des familles

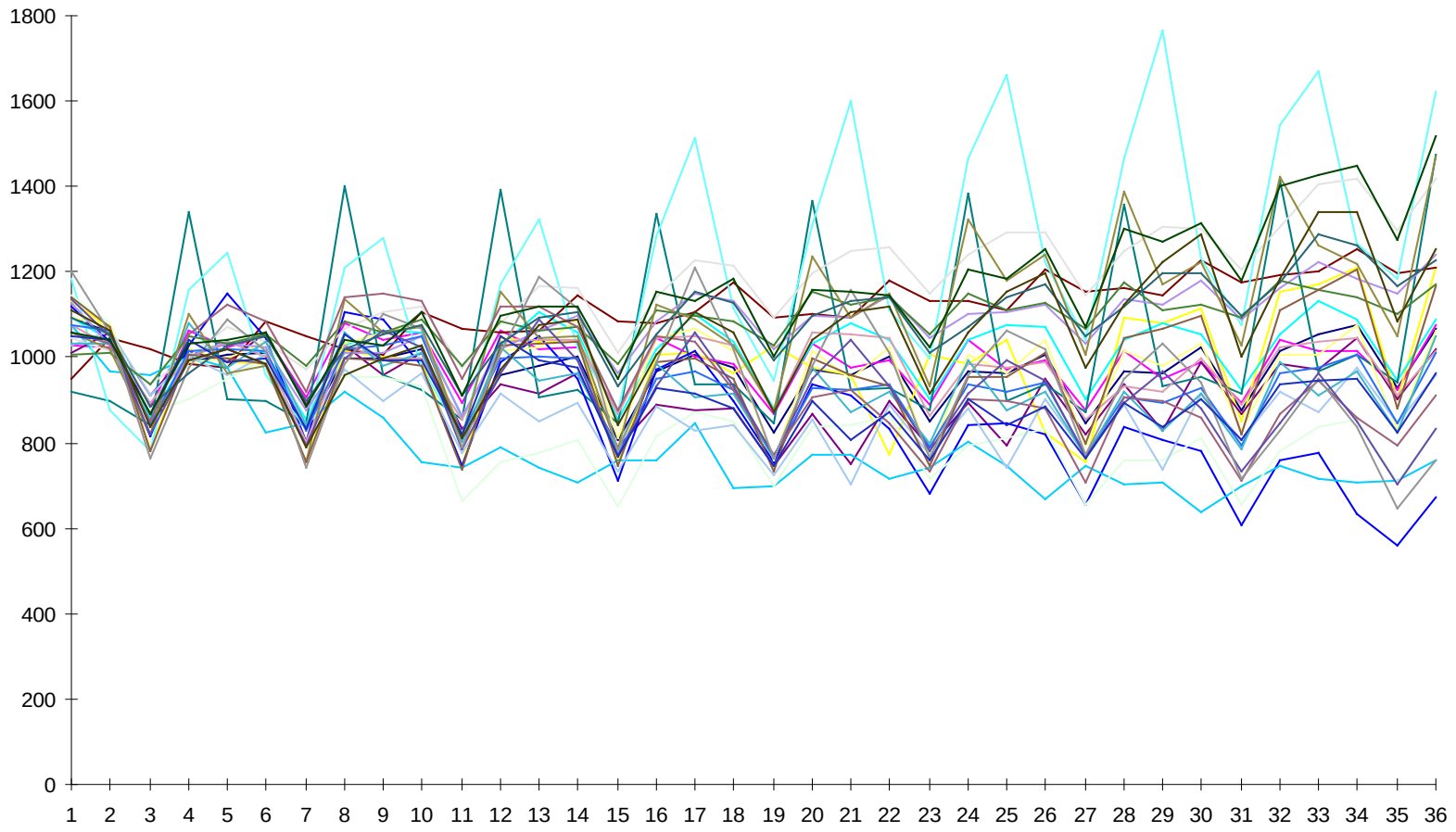


size effect



speed

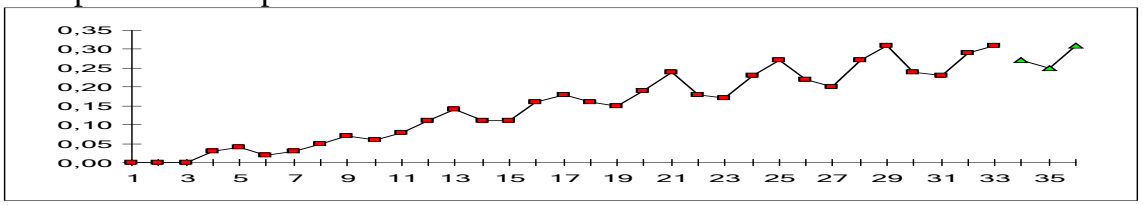
LA PRODUCTION INDUSTRIELLE EN FRANCE DE 1980 A 1988



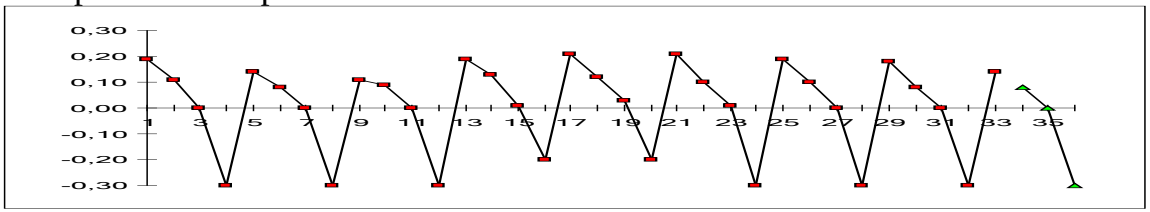
From G.Cohen's Ph.D, 1999

LES SIX PREMIÈRES COMPOSANTES TEMPORELLES ET LEURS PRÉVISIONS

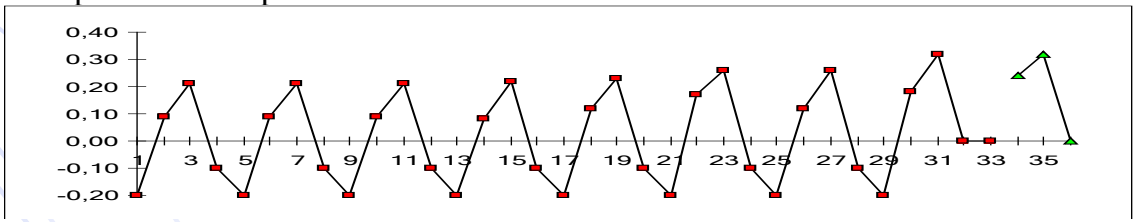
Composantes temporelles N°1



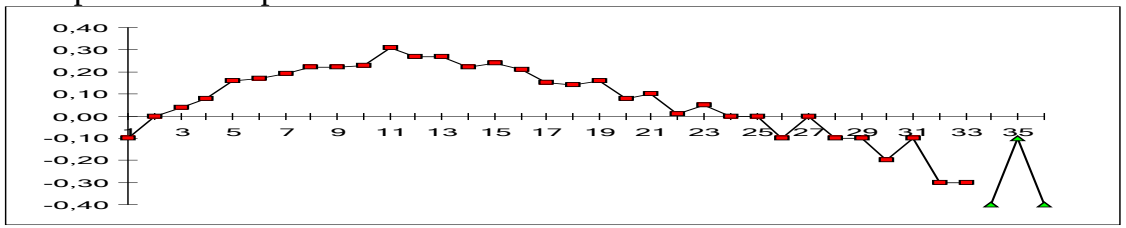
Composantes temporelles N°2



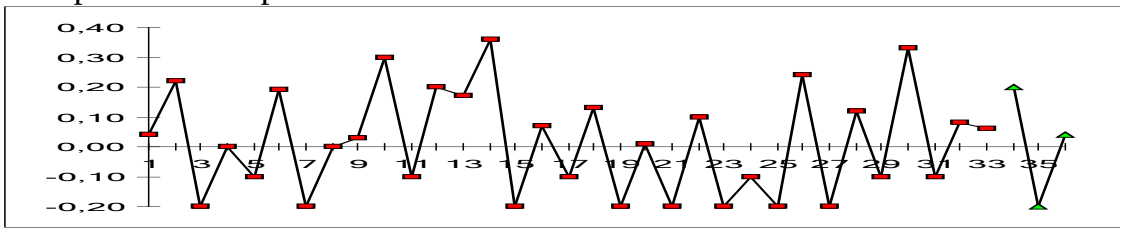
Composantes temporelles N°3



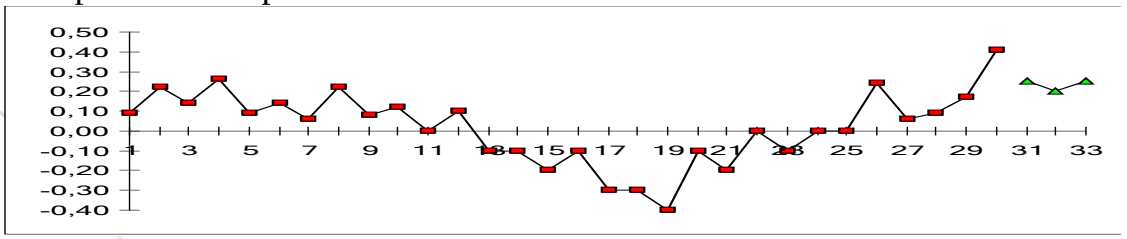
Composantes temporelles N°4



Composantes temporelles N°5



Composantes temporelles N°6



I.4 Résolution numérique

Solve either: $\int_0^T C(t,s) f(s) ds = \lambda f(t)$ or $\frac{1}{n} W \xi = \lambda \xi$

Analytical solutions for some well – known process (brownian motion, brownian bridge)

For a sample of trajectories:

W matrix n × n (multidimensional scaling);

$$w_{ij} = \int_0^T x_i(t)\, x_j(t) dt$$

w_{ij} may be computed for jump processes ⇒ exact solutions ⇒

$$f(t) = \frac{1}{n} \frac{1}{\lambda} \sum_{i=1}^{n} \xi_i X_i(t)$$

But not feasible for large n

Discretisation $t_0, t_1, \ldots, t_p$

Numerical integration of $\int_0^T C(t,s)\, f(s) ds = \lambda f(t)$

$$\sum_{t=0}^{p} C(t,t_j)\, f(t_j) a_j = \lambda f(t) \qquad CAf = \lambda f(t)$$

A diagonal weight matrix

- ✓Rectangular integration: $a_j = t_{j+1} - t_j$
- ✓Trapezoidal integration: $a_0 = \frac{t_1 - t_0}{2}$ $a_j = \frac{t_{j+1} - t_{j-1}}{2}$ $\ldots a_p = \frac{t_p - t_{p-1}}{2}$
- ✓ Simpson …

II. Analyse des correspondances fonctionnelle

Ou « analyse harmonique qualitative » (Deville, Saporta 1979)

- Évolution d'une variable qualitative
  « Event – history data »

- Trajectoires d'un processus qualitatif

  - Etats matrimoniaux
  - Phases du sommeil
  - Mobilité géographique

- Cas général

$X_t \qquad t \in [0; T]$

Principe barycentrique en temps continu:

$$\begin{cases} z = \alpha \frac{1}{T} \int_0^T X_t a_t dt \\ z_i = \alpha \frac{1}{T} \int_0^T \sum_x a_t^x 1_t^x (i)\, dt \end{cases}$$

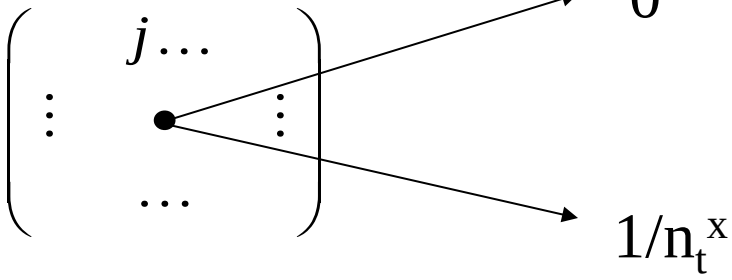
$$a_t = \alpha N_{tt}^{-1} X_t^{'} z \qquad a_t = \left(a_t^1, ..., a_t^n \right)$$

$$\frac{1}{T} \int_0^T N_{tt}^{-1} N_{ts} a_s ds = \lambda a_t \text{ équation intégrale}$$

$$\frac{1}{T} \left[\int_0^T A_t dt \right] z = \lambda z \text{ équation matricielle}$$

- « Multidimensional scaling » avec un indice de présence – rareté

$$A_t = X_t \left(X_t' X_t \right)^{-1} X_t' = i$$



A_t – matrice de similarité

$\int A_t dt$ matrice intégrée, positive définie

~ produits scalaires

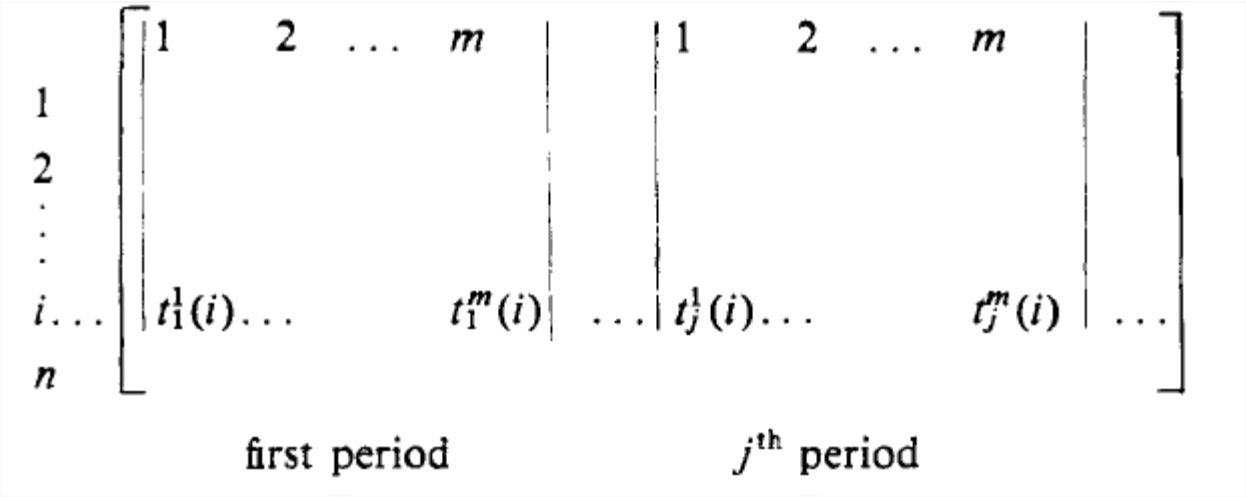
↓

représentation euclidienne

- Résolution numérique
  - ✓Exacte

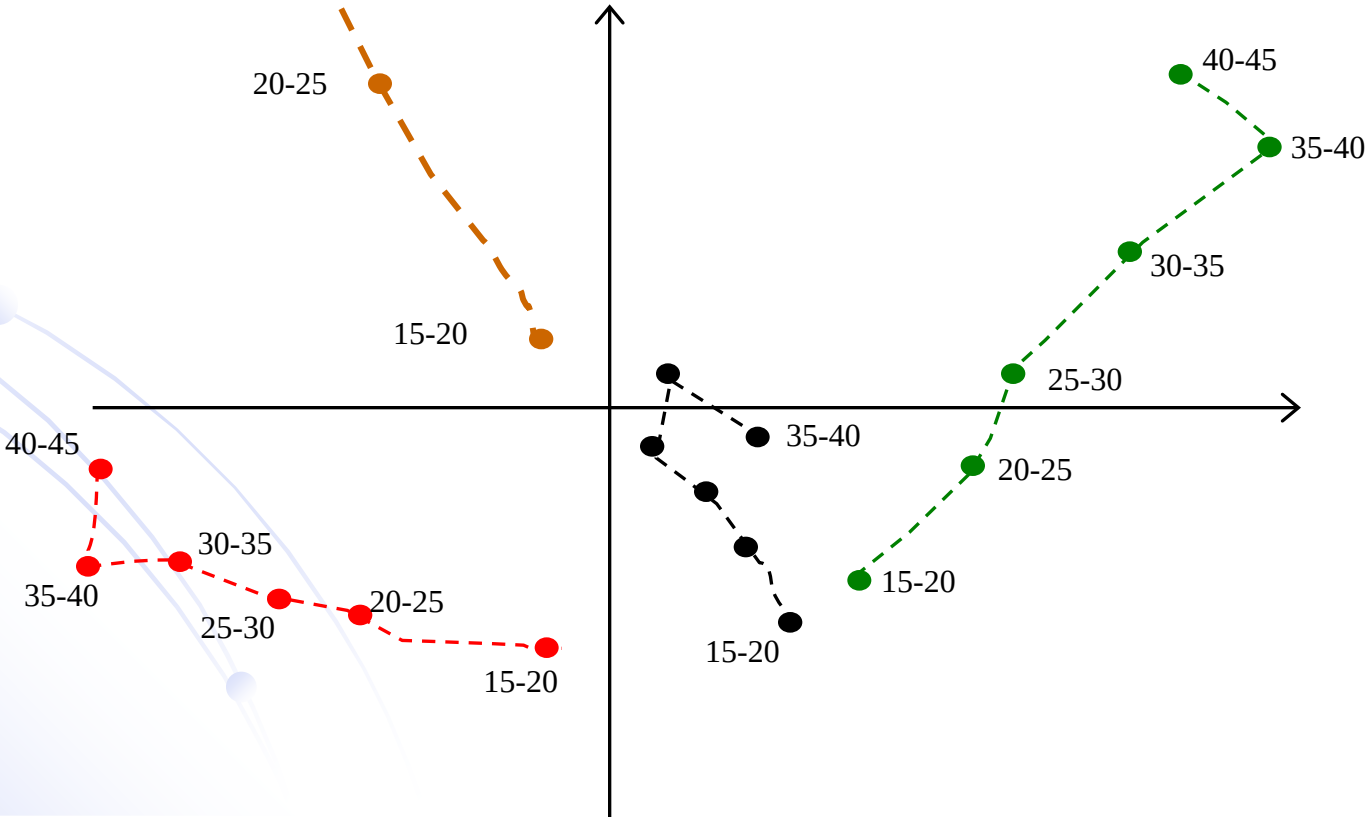
  Rassemblement de <u>toutes</u> les dates de changement d'état

- Approchée
  - ✓Décomposer T en p périodes 0, t_1,…, T
  - ✓Approximation par des fonctions constantes par intervalles
  - ✓ AFC du tableau contenant les temps passés par un individu i dans chaque état et selon les intervalles t_j, t_{j-1}

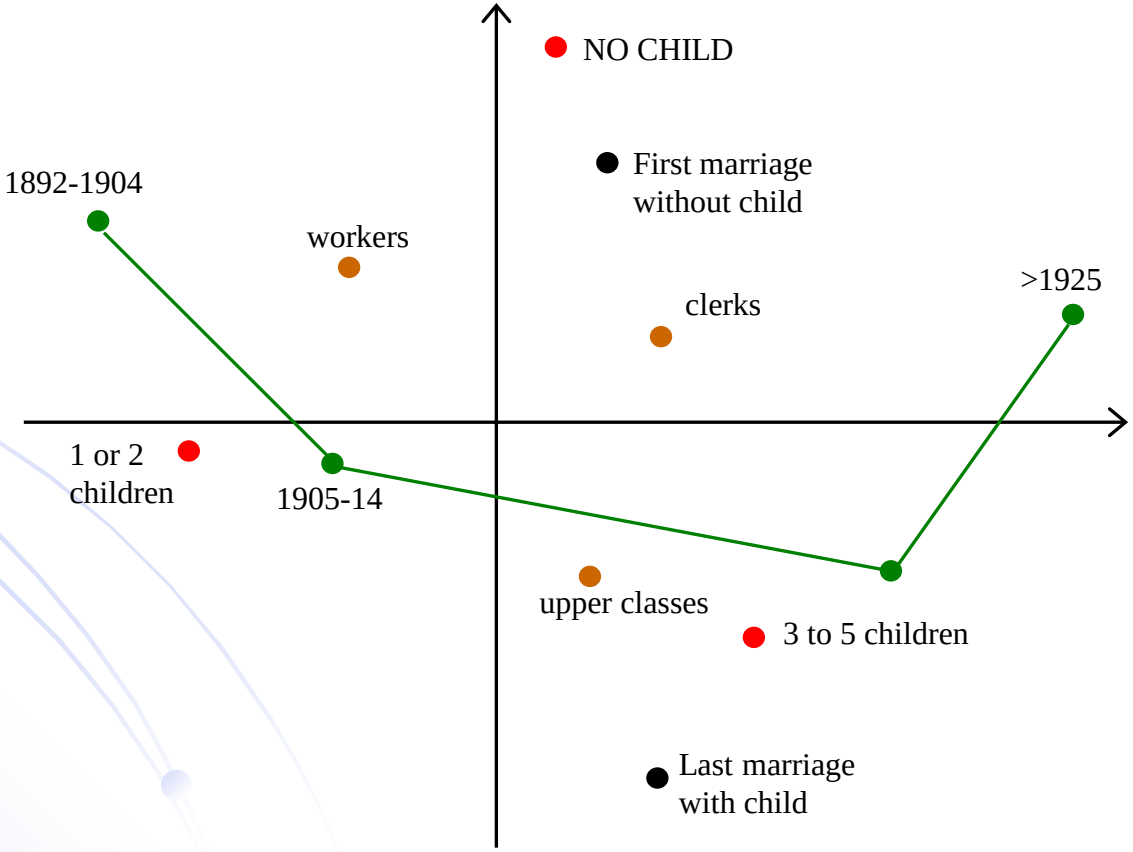


•Exemple (Deville 82):
French women married more than 3 times

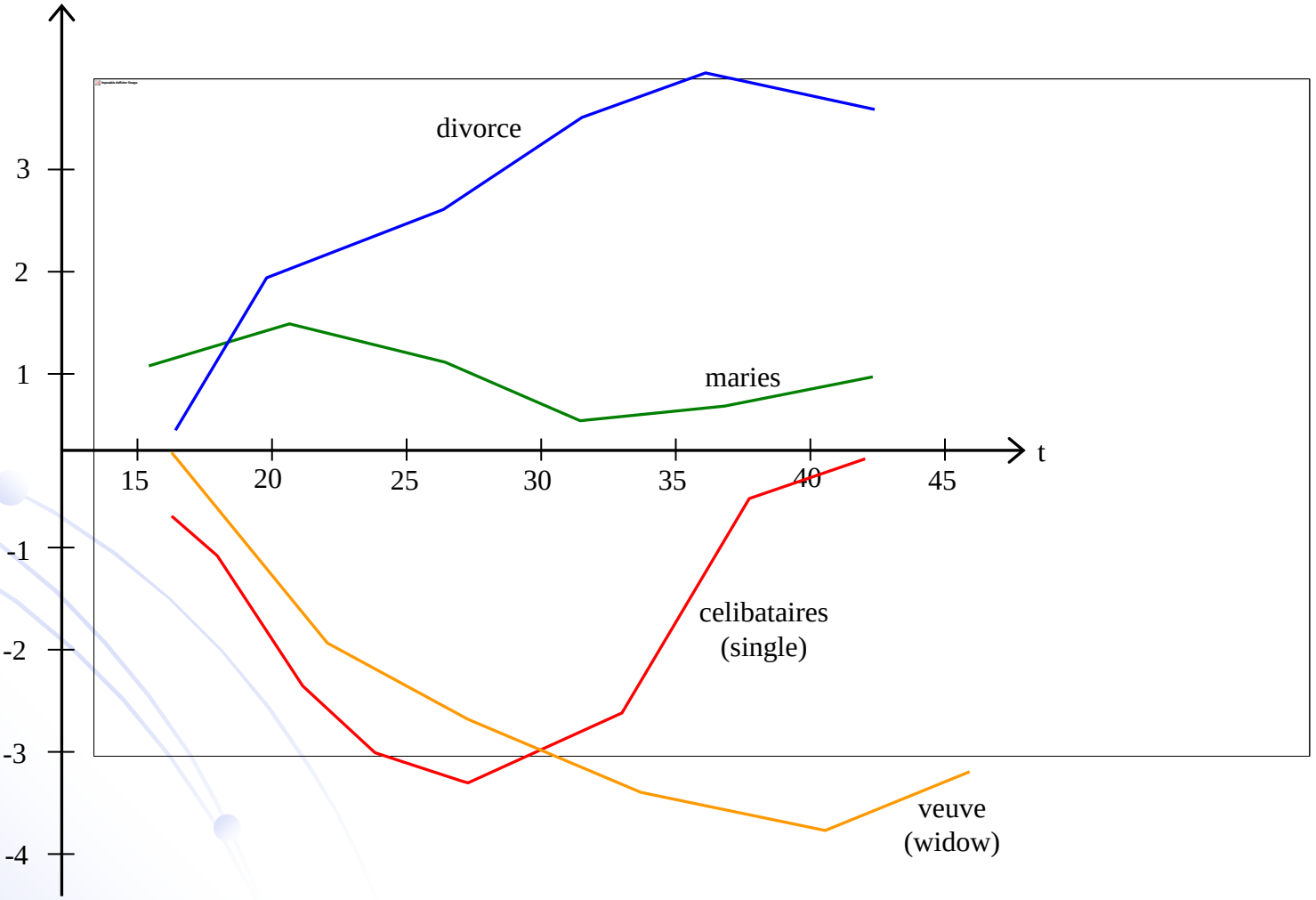
n=423 m=4 single – – – t∈[15; 45]
married – – –
divorced – – –
widowed – – –



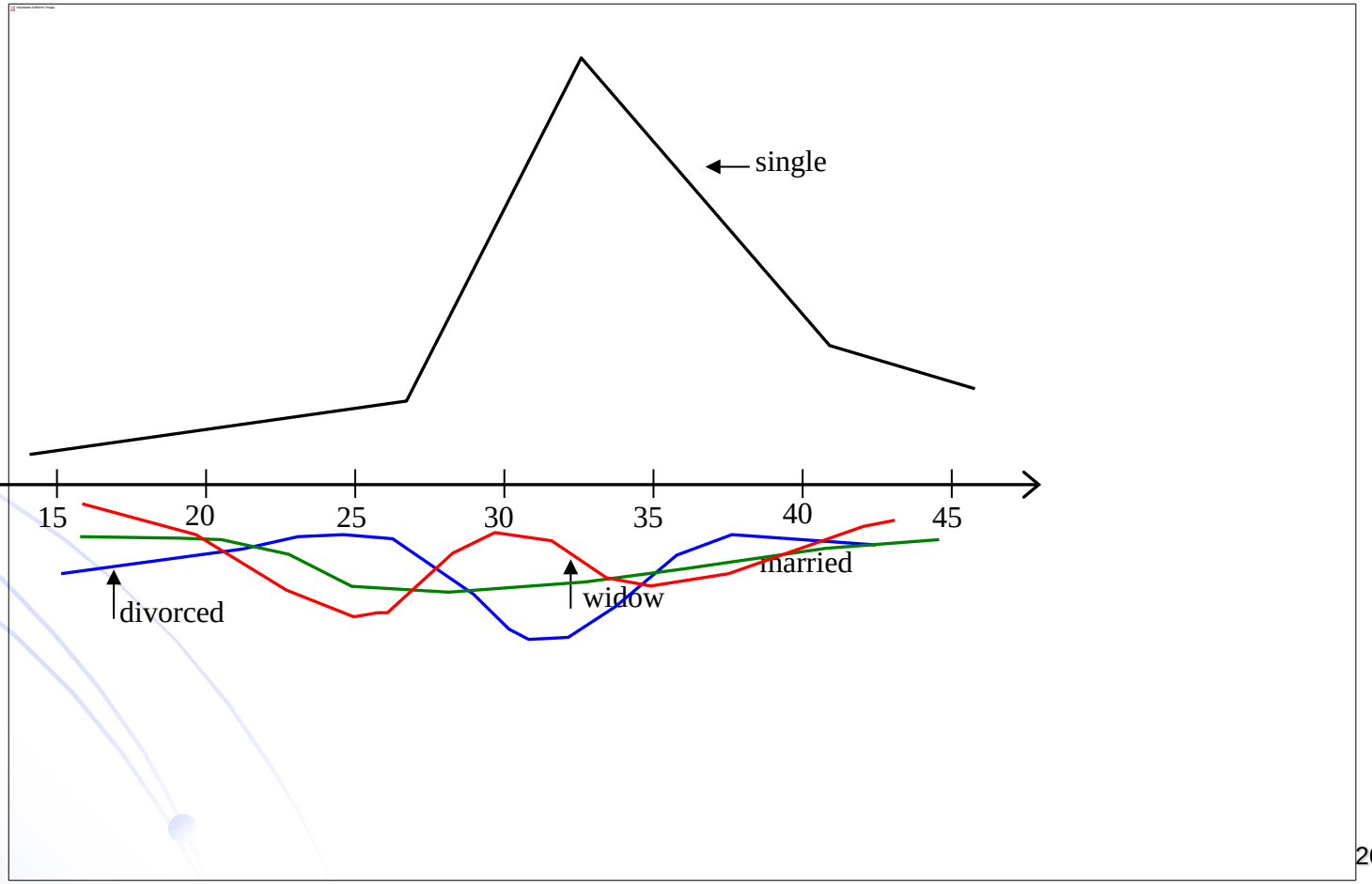
•Variables supplémentaires



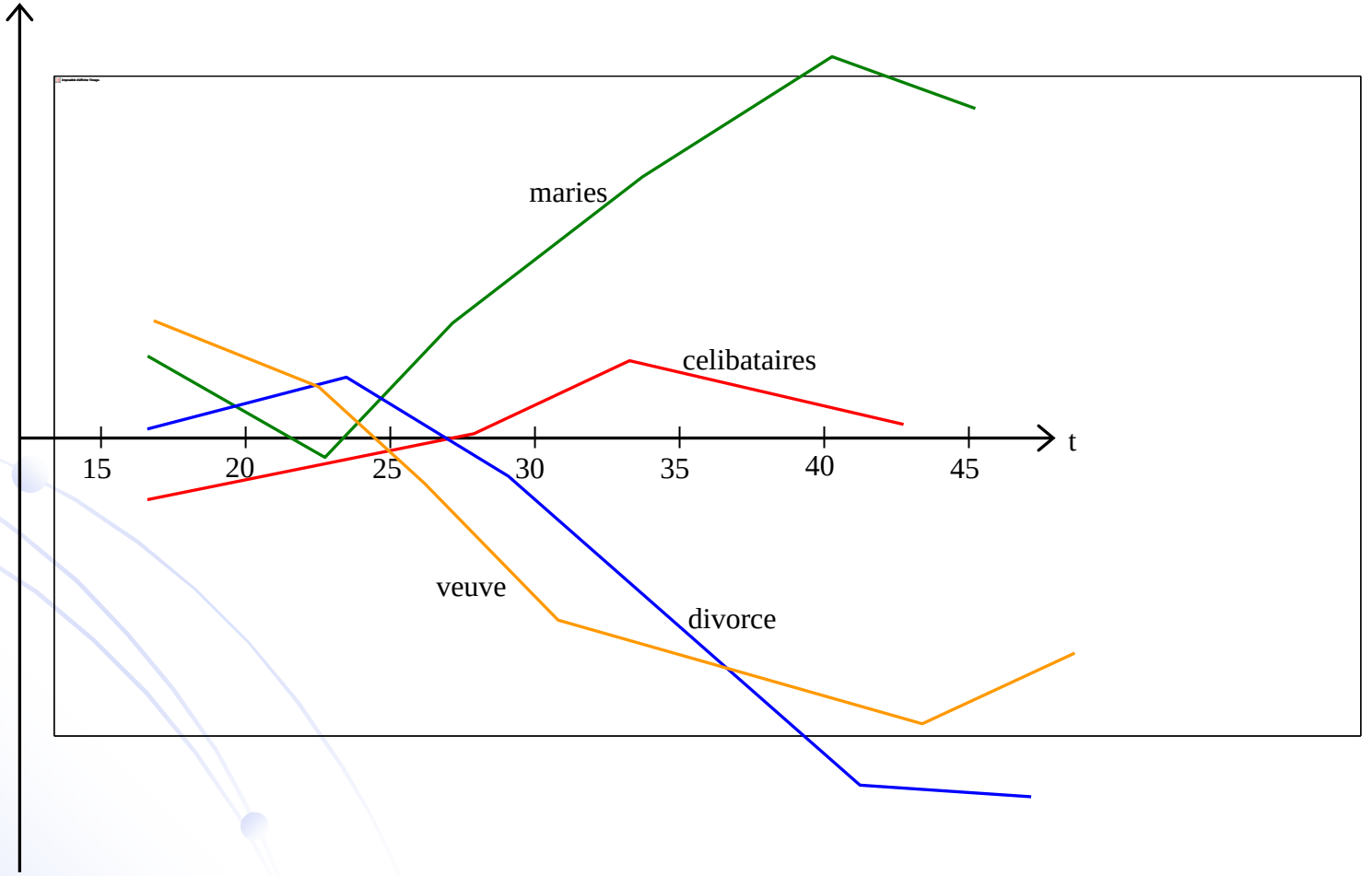
Premier facteur



- Deuxième facteur



- Troisieme facteur



- Conclusions
  - Analyse exploratoire de trajectoires par généralisation de l'ACP et de l'AFC
  - Possibilité de classification à l'aide des coordonnées factorielles
  - Autres approches TraMineR : http://mephisto.unige.ch/traminer/

References

- COHEN G. , 1999 *Contribution à la prévision des processus aléatoires par l'analyse harmonique*, Ph.D. CNAM
- DEVILLE J.C., 1974, « Méthodes statistiques et numériques de l'analyse harmonique », *Annales de l'INSEE* 15, 3-101
- DEVILLE J.C., SAPORTA G., 1979, « Analyse harmonique qualitative », *Data Analysis and Informatics,* E. Diday eds., North-Holland, 375-389
- DEVILLE J.C., SAPORTA G., 1983, « Correspondence analysis, with an extension towards nominal time-series », *Journal of Econometrics* 22, 169-189
- HEIJDEN PGM van der.,1987, *Correspondence analysis of longitudinal categorical data*, DSWO Press, Leiden
- PREDA C. , 1999, *Analyse factorielle d'un processus*, Ph.D. Université Lille 1
- RAMSAY, J.O. and SILVERMAN, B.W. ,2005: *Functional Data Analysis.* 2nd ed. Springer
- RAMSAY, J.O. and SILVERMAN, B.W., 2002: *Applied Functional Data Analysis. Methods and Case Studies*. Springer
- SAPORTA G., 1985, « Data analysis for numerical and categorical individual time-series », *Applied Stochastic Models and Data Analysis* vol.1., n°2, 109-119