Statistique et analyse des données

Jean-Claude Deville

Analyse et prévision des séries chronologiques multiples non stationnaires

Statistique et analyse des données, tome 3, n° 3 (1978), p. 19-29

<http://www.numdam.org/item?id=SAD_1978__3_3_19_0>

Statistique et Analyse des Données
3 - 1978 pp. 19, 29

ANALYSE ET PREVISION DES SERIES CHRONOLOGIQUES MULTIPLES NON STATIONNAIRES

Jean-Claude DEVILLE
INSEE

1. UN PEU DE DEMOGRAPHIE

En 1962, l'INSEE a réalisé une enquête permettant de reconstituer le calendrier complet de descendance pour plus de 100 000 familles (date de mariage et date de naissance des enfants, au jour près). Une nouvelle enquête,réalisée en 1975 sur un échantillon de même taille, est en cours d'exploitation. Entre temps, il s'est passé bien des choses et pas mal d'encre a coulé à propos de la chute de la natalité.

Deux problèmes essentiels se posent: comment analyser le processus de constitution des familles et comment prévoir la descendance finale d'une promotion de mariage quand on ne connaît que le début de sa vie reproductive ?

Des techniques d'analyse de données sur les processus aléatoires permettent plus ou moins bien de résoudre ces problèmes. On se rendra compte, qu'en fait ce n'est pas seulement en démographie que cette démarche peut avoir son intérêt.

Pour lier ce qui précède à ce qui suit, disons que nous considérons le calendrier de constitution des familles comme un processus aléatoire $X_t(\omega)$ (nombre d'enfants de la famille ω au bout du temps t) du second ordre continu en moyenne quadratique.

2. PROCESSUS DU SECOND ORDRE ET VECTEUR ALEATOIRE HILBERTIEN

$(\Omega, \mathcal{A}, P)$ étant un espace probabilisé et T un intervalle compact de $\mathbb{R}$,muni de la tribu borélienne, un processus du second ordre est une application de T dans $L^2(\Omega, \mathcal{A}, P)$.

Il est continu en moyenne quadratique si cette fonction est continue pour la norme de L^2 ; la moyenne $M_t = EX_t$ est alors une fonction continue ainsi que la fonction de covariance $C(t,t')=\mathrm{Cov}\ (X_t,X_{t'})$.

Si on suppose de plus X_t mesurable, on a :

$$V= \int_T \mathrm{Var}\ X_t\ dt= \int_{T\times\Omega} (X_t(\omega) - M_t)^2 dt\ dP(\omega) < +\infty$$

et (presque toutes) les trajectoires du processus sont des fonctions de carré intégrable sur T. Le processus X_t définit donc un vecteur aléatoire X dans l'espace de Hilbert $H = L^2(T,dt)$.

On voit facilement que $EX = M$ et $V = E\,\|X-M\|^2$

L'opérateur de covariance C dans H est défini par :

$\forall u\ \forall v$ dans $H : (u|Cv) = E\ (u\,|\,X-M)(X-M|v)$,

ou encore :

$C = E\ (X-M) \otimes (X-M)$ (★)

L'opérateur C est hermitien, positif, compact de trace finie et $\mathrm{tr}\ C = E\,\|X-M\|^2 = V$. La relation entre cet opérateur et la fonction de covariance est :

$$Cu(t) = \int_T C(t,s)\ u(s)\ ds.$$

3. DECOMPOSITION DE KARUHENEN-LOEVE (analyse harmonique)

D'après ce qui précède C admet une représentation spectrale discrète :

$$C = \sum_{i=1}^{\infty} \lambda_i\ f_i \otimes f_i \qquad (1)$$

Dans cette formule :

- $\{ f_i \}$ est un système orthonormé de vecteurs propres de C, f_i ayant pour valeur propre λ_i .

(★) Rappelons que dans un espace de Hilbert on peut définir le produit tensoriel $u \otimes v$ comme pl'application linéaire de rang 1 : $x \longmapsto (x|u)v$.

Si u est unitaire $u \otimes u$ est alors le projecteur orthogonal sur la direction de u.

- $\{\lambda_i\}$ forme une suite décroissante de nombres positifs ayant pour limite 0. Dans (1) chaque valeur propre est répétée autant de fois que sa multiplicité. Les f_i sont déterminés (au signe près) de façon unique dès que λ_i est valeur propre simple.

Posons $\xi_i = \lambda_i^{-1/2} (X-M|f_i)$. On vérifie immédiatement que $E\,\xi_i = 0$ et que $E\,\xi_i\,\xi_j = 1$ si $i = j$, 0 si $i \neq j$; les ξ_i forment un système orthonormé de $L = L^2(\Omega, \mathcal{A}, P)$. On a en outre :

$$X-M = \sum_{i=1}^{\infty} \lambda_i^{1/2}\, \xi_i\, f_i \qquad (2)$$

ou sous forme fonctionnelle :

$$X_t(\omega)-M_t = \sum_{i=1}^{\infty} \lambda_i^{1/2}\, \xi_i(\omega)\, f_i(t) \qquad (2')$$

(développement de Karuhenen-Loeve)

Sous la forme (2) on a la généralisation des équations de l'analyse factorielle,(*) sous la forme (2') un développement canonique du processus aléatoire, les termes étant rangés par ordre d'importance. Toute la terminologie de l'analyse factorielle (facteurs, part de variance expliquée) se transporte donc dans le domaine des processus aléatoires. "L'analyse harmonique" est la statistique des processus du second ordre basée sur ce développement. Nous n'insisterons pas sur les problèmes purement statistiques (estimations des λ_i et f_i) ou numérique (approximation ou interpolation) qui se posent dans ce contexte.

4. ESPACES ET OPERATEURS

La famille (f_i) forme une base hilbertienne du sous espace $\overline{\mathrm{Im}C}$ de H. De même la famille ξ_i forme une base hilbertienne d'un sous-espace L(X) de L; ce sous espace n'est autre que l'espace engendré par les variables aléatoires centrées $X_t - M_t$. En effet, ξ_i appartient par définition même de l'intégrale de Riemann, à l'espace de Hilbert engendré par les variables $X_t - M_t$ de sorte que L(X) est contenu dans cet espace. Mais, d'après (2'), $X_t - M_t$ appartient à L(X) car $\Sigma_i [\lambda_i^{1/2} f_i(t)]^2 = C(t,t) < \infty$. L'espace engendré par les $X_t - M_t$ est donc contenu dans L(X).

(*) Pour une étude systématique des généralisations de l'analyse factorielle, entre autre aux processus, voir la thèse de POUSSE et DAUXOIS [1]

L'opérateur d'Escoufier U est défini sur $L^2(\Omega, \mathcal{A}, P)$ par (cf [4]:

$$U\xi = \int_T E\,[\ \xi(X_t - M_t)\ ]\ (X_t - M_t)\ dt.$$

En utilisant (2) on voit que :

$$U\xi = \sum_{i=1}^{\infty} \lambda_i\ E\ (\xi_i \xi\ )\xi_i \quad \text{soit} \quad U = \sum_{i=1}^{\infty} \lambda_i\ \ \xi_i \otimes \xi_i \quad .$$

Si on construit un isomorphisme A entre $L^2(\Omega,\ \ ,P)$ et H tel que pour tout i: $A\ \xi_i = f_i$, on voit que l'on a :

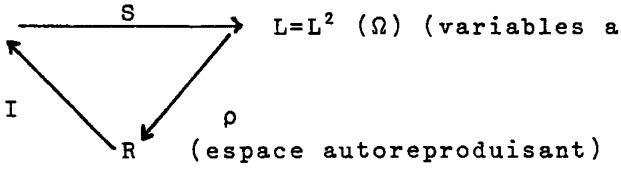
$$C = A\ U\ A^{-1}$$

Les opérateurs C et U sont "identifiés" par cet isomorphisme.

On peut compléter ce schéma en utilisant l'espace à noyau reproduisant associé au processus X_t (*). A toute variable ξ de L on associe la fonction $\rho\xi(t) = E\ \xi(X_t - M_t)$. On voit facilement que ρ est une bijection linéaire entre L(X) et un espace de fonctions continues sur T qu'on munit d'une structure hilbertienne en transportant par ρ la structure de L(X). On notera R cet espace (dit autoreproduisant). On notera alors I l'injection canonique de R dans H. On notera enfin S l'application linéaire de H dans L définie par :

$$\forall f \in H = L^2\ (T)\ : \qquad Sf\ = \int_T f(t)\ (X_t - M_t)\ dt.$$

Nous avons donc construit le diagramme suivant :



Il est clair qu'on a alors : $C = I \circ \rho \circ s$

et $U = S \circ I \circ \rho$

Pour compléter ce schéma, notons que l'opérateur $\Gamma = \rho \circ S \circ I$ n'est autre que la restriction à R de C. L'espace possède la propriété "autoreproductrice" suivante: soit c_t la fonction $C(t,.) = \rho\ X_t$. On a pour toute fonction φ de R :

$$(C\ \varphi)(t) = (c_t \mid \varphi\ )_R = \mathrm{Cov}\ (X_t, \rho^{-1}\varphi) = \varphi(t).$$

En particulier $(c_t \mid c_s)_R = C(t,s)$ et les fonctions c_t jouent dans R le rôle des masses de Dirac δ_t . Quant à l'espace R lui-même, son rôle, en statistique des processus du second ordre, est analogue à celui joué par la métrique de Mahalanobis en analyse multivariée.

(*) Une étude complète de ces espaces se trouve dans NEVEU [5]

On peut voir enfin qu'il y a correspondance entre les systèmes orthonormés (f_i) de H, (ξ_i) de L et $(\lambda_i^{1/2} f_i)$ de R de la façon suivante :

$$Sf_i = \lambda_i^{1/2} \xi_i \quad (\| Sf_i \| = \lambda_i^{1/2})$$
$$\rho\xi_i = \lambda_i^{1/2} f_i \quad (\| S\xi_i \| = 1)$$
$$I(\lambda_i^{1/2} f_i) = \lambda_i^{1/2} f_i \quad (\| I(\lambda_i^{1/2} f_i) \| = \lambda_i^{1/2}).$$

5. APPLICATIONS DE L'ANALYSE HARMONIQUE A LA PREVISION

On désire prévoir au mieux la variable X_t connaissant le processus sur l'intervalle $[o,T_o]$ $(T_o < T)$. On suppose par commodité que toutes les variables aléatoires sont centrées . Théoriquement le problème est simple: il suffit de prendre la projection de X_t sur l'espace L_o engendré par les variables X_t $(o \leq t \leq T_o)$(*). On voit aussi facilement que l'image dans R de cette projection n'est autre que ρX_T . En utilisant la base ξ_i fournie par l'analyse harmonique du processus sur $[o,T_o]$, on obtient donc le prédicteur :

$$\hat{X}_T = \sum_{i=i}^{\infty} c_i \xi_i \qquad \text{avec } c_i = EX_T \xi_i$$

On obtient un prédicteur approché en tronquant cette série au bout de k termes, de façon à pouvoir mener à bien le calcul .

On pose donc : $\hat{X}_T^k = \sum_{i=1}^{k} c_i \xi_i$

En utilisant le fait que $\xi_i = \lambda_i^{1/2} \int_o^{To} X_t f_i(t)\, dt$ on trouve que :

$$\hat{X}_T^k = \int_o^{To} g_k(t) X_t \, dt \qquad (3)$$

avec $$g_k(t) = \sum_{i=1}^{k} \lambda_i^{1/2} c_i f_i(t) \qquad (4)$$

et $$c_i = \lambda_i^{1/2} \int_o^{To} C(T,t) f_i(t)\, dt \qquad (5)$$

Malheureusement, si la suite $\hat{X}_T^k$ converge vers $\hat{X}_T$, la suite des fonctions g_k ne converge généralement pas vers une "bonne fonction". En fait X_T n'admet pas nécessairement une représentation intégrale comme dans (3). Pour s'en convaincre il suffit de penser au cas d'un processus à accroissements indépendants (Poisson ou Wiener par exemple); dans ce cas on aura $\hat{X}_T = X_{To}$ et la suite g_k convergerait, dans un sens à définir, vers la masse de Dirac posée en T_o.

(*) On obtient ainsi la meilleure prédiction linéaire de X_T.

Néanmoins la forme (3) a l'avantage de ne tenir compte que des composantes fortement explicatives du processus jusqu'à T_o. Les composantes lointaines forment une espèce de bruit de fond qu'il est plus utile d'éliminer que de conserver.

Cette méthode peut s'améliorer en utilisant des variables "toutes faites" comme X_{To} par exemple. On mettra alors le prédicteur sous la forme :

$$\hat{X}_T = \alpha X_{T.} + \tilde{Y}$$

où $\tilde{Y}$ est la projection de X_T sur l'orthogonal de X_{To} dans L_o. On peut surtout améliorer en perdant un peu de linéarité pour se rapprocher d'une espérance conditionnelle. Nous ne développerons pas cette idée dans ce papier.

Ceci posé, la prédiction repose sur l'axiome bien connu: "Tout est dit depuis qu'il y a des hommes et qu'ils pensent !". On prédit l'avenir d'une trajectoire du processus en cherchant comment, dans le passé, des trajectoires commençant de façon identique ou analogue ont évoluée . S'il se passe un phénomène nouveau, si la liaison entre les événements antérieurs et postérieurs à T_o subit une modification, aucune précision n'est possible sinon par le marc de café ou la boule de cristal.

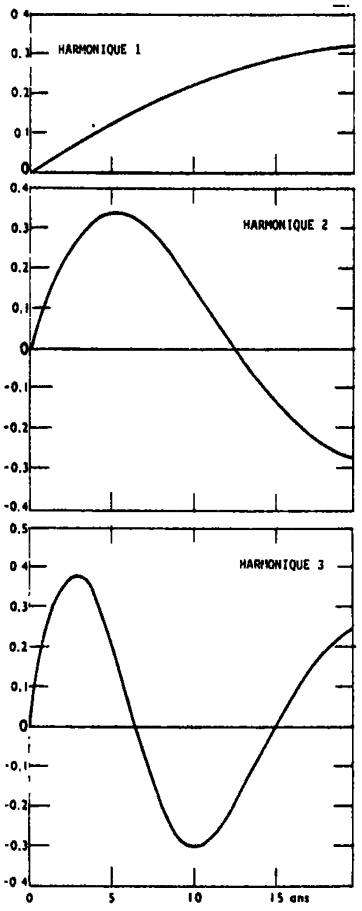
6. ENCORE UN PEU DE DEMOGRAPHIE

L'analyse harmonique de 98 785 familles suivies pendant 20 années ou de 128 515 familles suivies pendant 15 années a fourni les résultats suivants résumés par les graphiques 1 et 1' et le tableau 1.

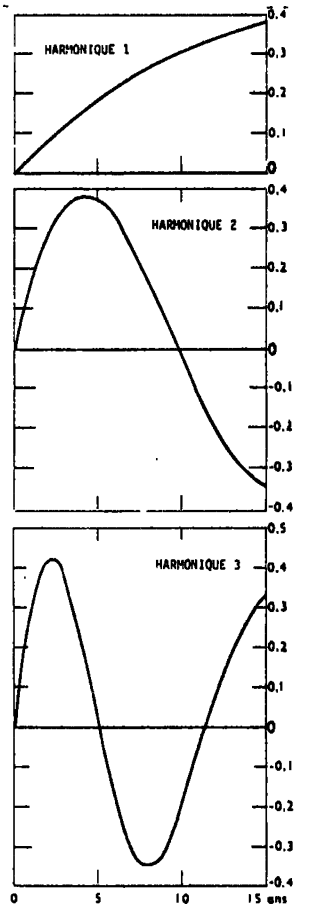
Le premier facteur est une mesure de la fécondité totale des familles et est très lié au nombre final d'enfants. Le second mesure la vitesse de constitution des familles. Le troisième, d'interprétation moins évidente donne une composante négative($\xi_3 < 0$) si les familles se constituent en un court intervalle de temps,si les naissances successives sont rapprochées.

On tire de cette interprétation des résultats assez intéressants sur le plan démographique. Nous signalerons seulement ici les plus notables d'entre-eux.

Le troisième axe est très lié au niveau culturel comme on le voit au graphique 2 où sont représentés les points moyens des différentes catégories socio-professionnelles. De ce fait, le graphe-plan 1-3 restitue la courbe classique en J renversé liant le niveau social à la fécondité. Il apparaît aussi que les catégories de travailleurs indépendants ont tendance à constituer plus rapidement (à stopper plus tôt le développement de) leur famille que les catégories de salariés.



Graphique 1 — Harmoniques du calendrier de constitution des familles sur 20 années de mariage (98 785 familles).

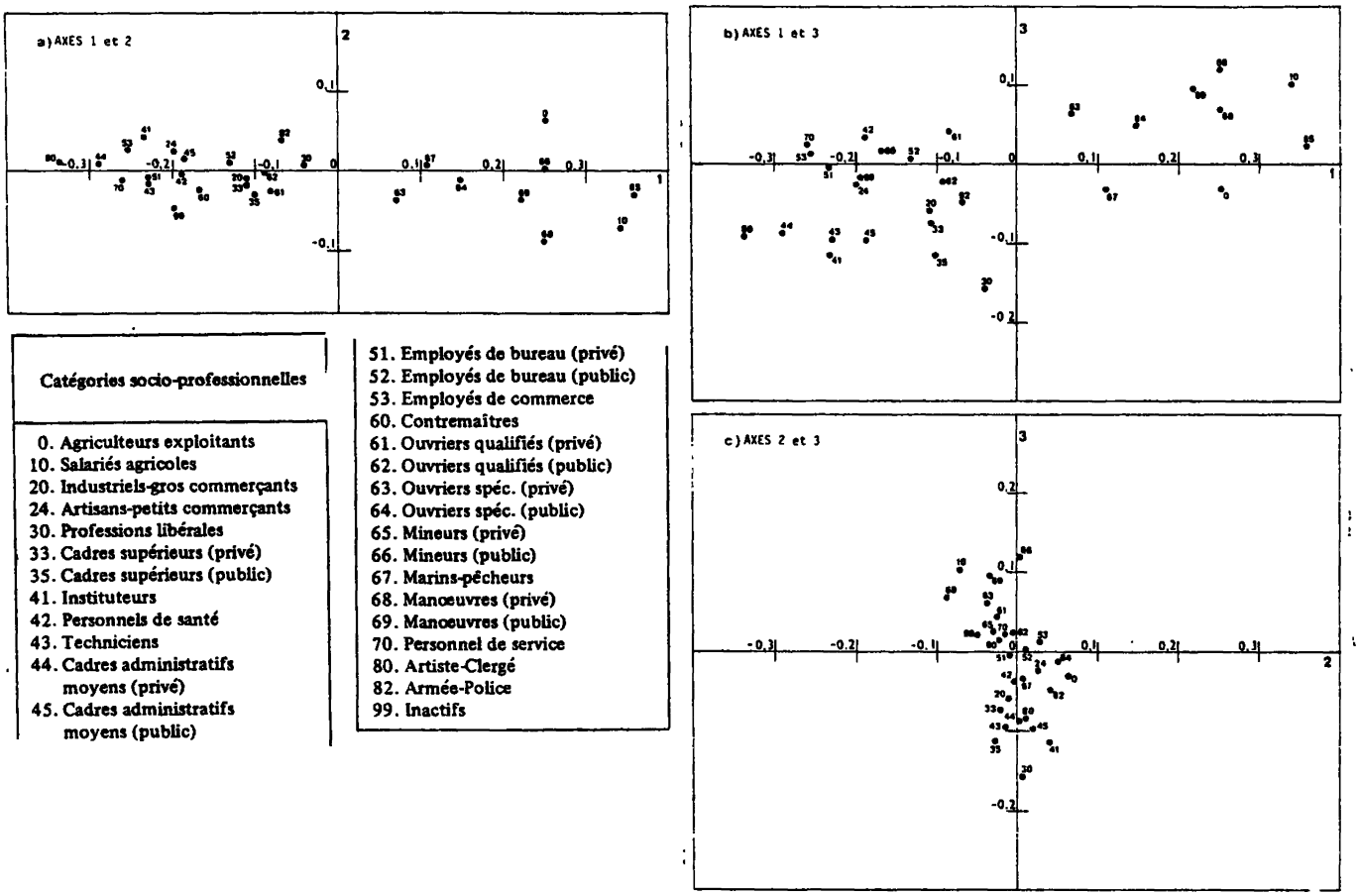


Graphique 1' — Harmoniques du calendrier de constitution des familles sur 15 années de mariage (128 515 familles).

TABLEAU 1 — PART DE VARIANCE EXPLIQUÉE PAR CHAQUE FACTEUR

Facteur n°	1	2	3	4	5	6	7	8	9	10
Etude sur 20 ans	92,83	4,05	1,01	0,52	0,36	0,21	0,14	0,12	0,11	0,09
Etude sur 15 ans	91,11	4,73	1,40	0,70	0,42	0,29	0,21	0,16	0,13	0,11

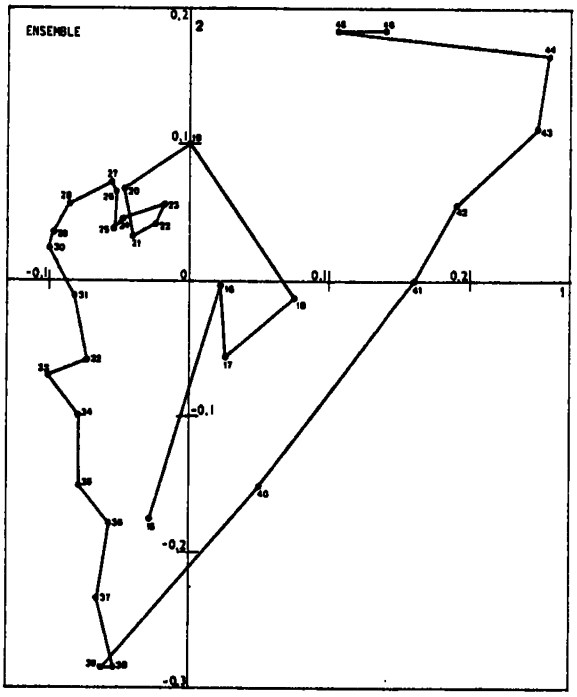
On constate aussi (graphique 3) que l'évolution démographique française des années 20 aux années 60 a été plus complexe qu'on ne le pensait. En particulier les années 30 ont été caractérisées par un retard de plus en plus grand dans la constitution des familles. Comme la descendance finale n'a que peu varié pour les mariages conclus à cette époque, on en conclut que la "catastrophe" démographique des années 30 a été très fortement surestimée.



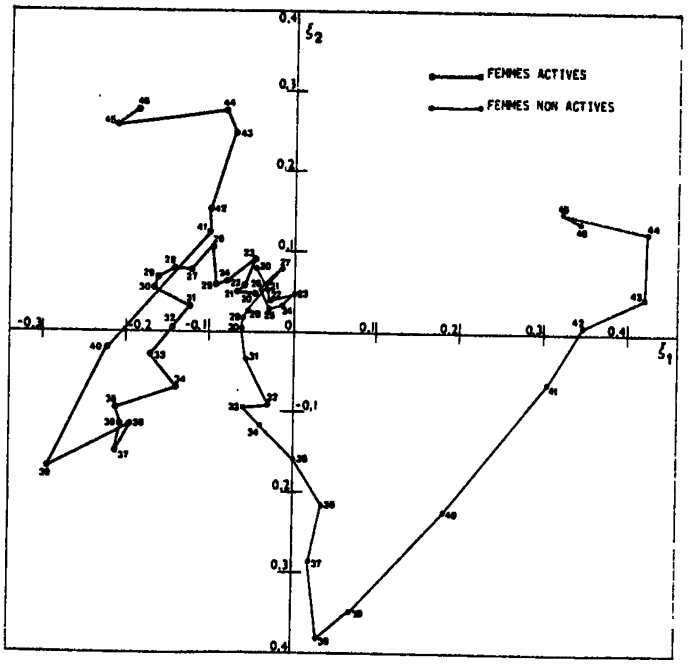
Graphique 2. — Catégories socio-professionnelles dans le plan des axes factoriels.

Enfin, (graphique 4), il apparait que cette évolution a été très différente selon que les femmes étaient actives ou pas. Schématiquement, les inactives avaient plus d'enfants après la guerre qu'avant mais au même rythme; par contre, les actives avaient un peu moins d'enfants mais constituaient plus rapidement leur famille. Ces deux mouvements ont pour effet un accroissement de la natalité du moment et expliquent aussi une bonne part de la forte fécondité française au cours des années 1945 à 1965.

On a, par ailleurs, utilisé la technique de prévision linéaire sur un échantillon de 314 femmes; on a essayé de prédire la descendance au bout de 20 années en utilisant uniquement l'information donnée par les dix premières années de mariage.



Graphique 3 — Promotions de mariages dans le plan des deux premiers facteurs. Ensemble.



Graphique 4 — Femmes actives et femmes non actives. Paramètres ξ_1, ξ_2 annuels de 1920 à 1946.

Les résultats peuvent se résumer par le tableau suivant. On a classé les familles selon le nombre des enfants nés au bout de 10 années de mariage.

Nombre d'enfants au bout de 10 années de mariage	Nombre d'enfants moyens au bout de 20 années de mariage	Prévision	Nombre de familles
0	0,17	0,22	35
1	1,25	1,37	57
2	2,37	2,52	78
3	3,39	3,66	66
4	4,50	4,91	44
5	5,74	5,89	19
6	6,75	7,30	12
7	8,00	8,04	1
8	9,00	8,95	22

On constate que la prédiction surestime systématiquement la réalisation.La cause de ce phénomène résulte de l'utilisation d'une méthode linéaire de prédiction. Il est bien connu que la meilleure prédiction de X_T est l'espérance mathématique de X_T conditionnellement à la tribu des événements antérieurs à T_o . Cette variable coïncide avec le prédicteur linéaire uniquement dans le cas de processus gaussiens.

Pour améliorer la prédiction il faudrait donc abandonner un peu de linéarité pour rapprocher le prédicteur d'une espérance conditionnelle. On pourra par exemple utiliser une méthode linéaire à nombre d'enfants au temps T_o fixé; au lieu d'une analyse harmonique on en utilisera une demi-douzaine. Des travaux seront menés dans cette voie pour tenter d'analyser l'évolution récente de la fécondité à partir des données de l'enquête sur les familles de 1975.

Il ne faut pas oependant nourrir une espérance exagérée sur ce sujet. On a essayé, par exemple, de prédire dans les mêmes conditions que ce qui vient d'être expliqué, la descendance finale des familles à partir du nombre d'enfants nés au bout de 10 ans et de l'âge du plus petit. La qualité des résultats obtenus est comparable à ce qui vient d'être exposé (surestimetion systématique) et cela au prix d'une complication considérable des calculs numériques.

R E F E R E N C E S

[1] A.POUSSE et J.DAUXOIS — Les analyses factorielles en calcul des probabilités et en statistique: Essai d'étude synthétique Pub.du Labo. de statistique de l'Université Paul Sabatier, Toulouse-1976.

[2] JC.DEVILLE — Méthodes statistiques et numériques de l'analyse harmonique. Annales de l'INSEE n°15 - 1974.

[3] JC.DEVILLE — Analyse harmonique du calendrier de constitution des familles - POPULATION n° 1 - 1977.

[4] Y.ESCOUFIER — Echantillonnage dans une population de variables aléatoires réelles. Publication de l'ISUP vol.XIX 1970

[5] J. NEVEU — Processus aléatoires Gaussiens - Presse de l'Uniservité de Montréal - 1968.

[6] E. PARZEN — Time-séries analysis - HOLDEN-DAY -1967.