REVUE DE STATISTIQUE APPLIQUÉE

F. DOUSSET
R. CASAMITJANA
J. L. GROBOILLOT
A. WARNIER DE WAILLY

Estimation d'une matrice de Markov

Revue de statistique appliquée, tome 15, n° 1 (1967), p. 87-94
<http://www.numdam.org/item?id=RSA_1967__15_1_87_0>

ESTIMATION D'UNE MATRICE DE MARKOV*

F. DOUSSET et R. CASAMITJANA
(C.E.R.E.N.)

J. L. GROBOILLOT et A. WARNIER DE WAILLY
(FRANCORELAB)

Les méthodes qui cherchent à évaluer directement les consommations futures (à moyen ou long terme) de combustibles par les ménages se sont souvent révélées imprécisés lorsque l'on cherche à distinguer les diverses formes d'énergie. Toute consommation d'énergie se fait par l'intermédiaire d'un parc d'appareils (chaudières, cuisinières, poëles, radiateurs, réfrigérateurs...) dont l'évolution se traduit par des substitutions entre charbon, fuel, gaz et électricité. Il apparaît ainsi qu'il convient d'aborder de telles prévisions par le biais d'une étude du parc. De façon analogue, on peut penser baser l'étude de la consommation de carburants sur celle du parc de véhicules. Ce type d'approche présente cependant l'inconvénient de nécessiter une information relativement riche car elle doit porter non seulement sur l'état du parc, mais aussi sur les consommations unitaires d'énergie à y associer pour pouvoir passer aux consommations totales. Par contre, il permet de tenir compte de certaines variables exogènes dont l'influence est déterminante telles que le revenu, le type et l'âge de l'immeuble.

Le problème dont la résolution est évoquée dans les pages qui suivent s'inscrit dans le cadre d'une étude de ce type portant sur la prévision de l'équipement 1970 des ménages en appareils de chauffage. Une étude préalable ayant montré qu'il était nécessaire de séparer les logements suivant qu'ils étaient ou non de construction récente (postérieurs à 1954) (1) on a été conduits à mettre au point des méthodes d'approche appropriées à chacun des deux cas précédents. Ainsi pour l'étude des logements anciens, une méthode consistait à répartir les ménages en classes homogènes du point de vue "équipement chauffage". Cette répartition est connue à différentes dates grâce à des enquêtes par sondage sur l'équipement des ménages.

Un modèle représentatif de l'évolution des différentes classes choisies dans le temps a été mis en oeuvre. Ce modèle suppose que l'on à affaire à une chaine de Markov dont les probabilités concernent le passage d'une classe d'équipement à une autre. Ainsi par exemple est défini la probabilité pour un ménage muni de radiateurs ou de poëles de s'équiper d'un chauffage central fonctionnant au gaz et au fuel (2) (3).

---

* Article reçu en Juillet 1966.

(1) A titre d'illustration on peut citer que 14 % de l'ensemble des logements construits avant 1954 et existants en 1962 sont équipés du chauffage central, alors que 52 % de ceux construits postérieurement le sont.

(2) Un tel modèle a déjà été utilisé pour les équipements fonctionnant au charbon mais l'insuffisance des données n'avait pas permis de calculer la matrice des probabilités de passage sur ordinateur. M. E. ROSENFELD et M. SALOMON - "Utilisation de modèles markoviens dans les études du Marché" Bulletin de l'Institut International de Statistique - 1961.

(3) L'ensemble de cette étude a fait l'objet d'une communication de MM. DOUSSET et CASAMITJANA au Congrès de la Société d'Econométrie - Varsovie 1966 - sous le titre "Utilisation d'un modèle à processus stochastique pour la prévision de l'équipement des ménages en appareils de chauffage.

L'information statistique disponible permettant de connaître trois vecteurs effectifs comportant chacun dix classes d'équipement a fait apparaître une hiérarchie entre types d'équipement considérés ; Ainsi les ménages ne changent d'équipement que pour adopter un équipement d'ordre supérieur. Cette hiérarchie a permis de triangulariser la matrice des probabilités de passage.

L'estimation de la matrice de Markov à partir des données disponibles a nécessité la mise au point d'une méthode de calcul appropriée dont l'analyse fait l'objet du présent article.

Soit un système qui évolue suivant un processus de Markov discret, de matrice de passage annuelle A(P. P.) (entrée en ligne, sortie en colonne).

On connait la composition du système à différentes dates, c'est-à-dire le nombre d'individus dans chaque état à ces instants. Soient V_{t_1}, $V_{t_1 + \tau_1}$; V_{t_2}, $V_{t_2 + \tau_2}$; V_{t_3}, $V_{t_3 + \tau_3}$ vecteurs effectifs.

Les liaisons entre ces divers vecteurs de p composantes, définissant la composition du système à p états, sont exprimées comme suit :

$$V_{t_1} \xrightarrow{A^{\tau_1}} V_{t_1 + \tau_1}$$

$$V_{t_2} \xrightarrow{A^{\tau_2}} V_{t_2 + \tau_2}$$

$$V_{t_3} \xrightarrow{A^{\tau_3}} V_{t_3 + \tau_3}$$

Pour des raisons de collecte de l'information, on ne peut que considérer les trois passages indiqués.

On sait de plus, que la matrice de passage A est triangulaire inférieure (avec diagonale).

1 - ESTIMATION DE LA MATRICE DES PROBABILITES DE PASSAGE.

On se propose d'estimer statistiquement la matrice A. La méthode du maximum de vraisemblance semble dans ce cas délicate à appliquer et en conséquence on s'orientera vers un outil inspiré de la méthode des moments.

On notera, en effet, que :

$$\mathcal{E}[V'_{t_i + \tau_i}] = V'_{t_i} A^{\tau_i} \qquad (i = 1, \ldots\ldots 3)$$

(V' = transposé du vecteur colonne V)

($\mathcal{E}$ opérateur d'espérance mathématique)

On cherchera donc une matrice $\hat{A}$, telle que :

$$\mathcal{E}[V_{t_i + \tau_i}] = V_{t_i + \tau_i} \qquad (i = 1, \ldots\ldots 3)$$

soient les trois équations vectorielles :

$$V'_{t_i} \hat{A}^{\tau_i} = V'_{t_i + \tau_i} \qquad (i = 1, \ldots\ldots 3)$$

Ce système d'équations, sans doute, ne pourra être satisfait en raison des fluctuations d'échantillonnage.

Etant donné le contexte du processus, nous avons une approximation A_0 de la matrice $\hat{A}$. Nous posons :

$$\hat{A} = A_0 + Q$$

avec

$$Q = ||q_{ij}||$$
$$A_0 = ||a_{oij}||$$

Ce sont les éléments q_{ij} que nous allons chercher à estimer.

Pour cela, nous procèderons à une suite d'essais en donnant aux q_{ij} des valeurs prises "au hasard" selon une certaine fonction de répartition. Comme $\hat{A}_0$ est dès le départ une approximation de A, les tatonnements sur les q_{ij} seront de faible amplitude.

Les $q_{ij}(i \neq j)$ seront pris uniformément répartis sur le segment

$$\left[a_{oij}\left(1 - \frac{1}{3}\right) , a_{oij}\left(1 + \frac{1}{3}\right) \right] \quad \text{et}$$

q_{ii} est choisi de telle sorte que $\sum_{i=1}^{p} a_{ij} = 1$

La matrice Q est triangulaire inférieure comme $\hat{A}$.

1.1 - L'application de la méthode.

Pour chaque essai, on calcule les vecteurs W tels que :

$$V'_{t_i}(A_0 + Q)^{\tau_i} = W'_{t_i + \tau_i} \qquad (i = 1, \ldots\ldots 3)$$

et l'on compare les $W'_{t_i + \tau_i}$ calculés aux $V'_{t_i + \tau_i}$ observés. Si la distance est faible, nous pouvons dire que la matrice A ainsi trouvée est satisfaisante.

La distance de la solution trouvée à la solution cherchée peut être chiffrée par la somme des carrés des écarts* de $V_{t_i + \tau_i}$ et de $W_{t_i + \tau_i}$

$$S = \sum_{i=1}^{3} \sum_{j=1}^{p} (V_{t_i + \tau_i^j} - W_{t_i + \tau_i^j})$$

Si le S trouvé est inférieur au S calculé lors d'un essai antérieur on considère que la matrice A, ainsi calculée, est une meilleure estimation de la matrice $\hat{A}$ cherchée, et en conséquence, on la prendra comme nouvelle matrice A_0.

Mais si le S trouvé est supérieur au S correspondant aux essais antérieurs on considère l'essai comme nul, et on passe à l'essai suivant en conservant la précédante matrice A_0 comme meilleure estimation.

Lors des calculs, le nombre d'essais a été fixé à 1000, et dix matrices A différentes ont été rencontrées. La dernière matrice retenue était assurée d'être, parmi toutes celles qui avaient été essayées, celle qui convenait le mieux au modèle.

---

* $V^j_{t_i}$ est la jème composante du vecteur V_{t_i}

1.2 - Les contradictions de la méthode.

Le processus considéré est à évolution lente. Donc les termes de la diagonale principale de A sont voisins de 1, et les autres, s'ils ne sont pas nuls par hypothèse, sont voisins de 0.

On peut donc écrire A sous la forme :

$$A = I + E$$

où I est la matrice unité

et E une matrice triangulaire inférieure, donc les termes de la diagonale principale sont non nuls.

Alors, on a :

$$A^n \# I + nE$$

Ce qui prouve que pratiquement il y a indépendance entre les colonnes de la matrice E, chaque colonne est déterminée indépendamment des autres.

En effet, si

$$E = (e_1, e_2, \ldots\ldots, e_p)$$

et

$$V^j = (j^{ème} \text{ composante du vecteur } V)$$

et

$$G_i = \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} \longleftarrow i$$

on a :

$$V'_{t_i}(G_{j+\tau_i} e_j) = V^j_{t_i+\tau_i} \quad (i = 1, 2, 3) \qquad \text{(II)}$$

Les équations (II) peuvent être satisfaites de nombreuses façons si le nombre de composantes de e_j (qui est égal à P-J) est supérieur à 3. Mais, si le nombre de ces composantes est inférieur à 3, ces équations sont difficilement satisfaites. Le nombre de composantes de e_j sera d'autant plus faible que j sera grand.

Or, dans les résultats trouvés, les différences entre les composantes des vecteurs prévus (W) et les composantes des vecteurs observés (V), sont toujours de l'ordre de 0,5 %, sauf pour les composantes d'ordre supérieur(J élevé) où elles atteignent 2 %. Ces résultats sont en accord avec la remarque faite ci-dessus.

2 - LES PREVISIONS

La difficulté rencontrée lors de l'estimation de la matrice A nous a amenés à poser le problème différemment. La recherche de la matrice A n'est pas le but final du problème, mais le moyen de calculer la prévision pour 1970. En conséquence, nous avons envisagé d'effectuer le calcul de cette prévision pour toutes les matrices A qui ont été rencontrées successivement et retenues comme satisfaisantes, il apparaît que pour les 10 matrices conservées, chaque composante de la prévision a une étendue de 0,5 % excepté 3 composantes pour lesquelles l'étendue est de l'ordre de 1 %. (Ce sont les pourcentages des classes qui sont prévus).

Nous avons alors cherché à situer la variance des prévisions.

Soit un processus de Markov de matrice de passage annuelle A (P, P). Soit V le vecteur état du système au temps t = 0, c'est-à-dire

V^i = Pourcentage de ménage dans l'état i.

Au temps t = 1 l'effectif (en pourcentage) de l'état [1] a pour moyenne :

$$m = \sum_{i=1}^{p} a_{i1} V^i$$

et pour variance

$$\sigma^2 = \sum_{i=1}^{p} a_{i1}(1 - a_{i1}) V^i \qquad (A)$$

En effet, chaque individu de l'effectif V^i de l'état [i] ira dans l'état [1] avec la probabilité a_{i1}, on a là une variable binomiale et la variance du nombre d'individu venant de l'état [i] allant dans l'état [1] est en conséquence $V^i\ a_{i1}\ (1 - a_{i1})$. Les passages des individus de l'état [i] et des individus de l'état [j] dans l'état [1] étant deux phénomènes indépendants la variance σ^2 de l'effectif total arrivant dans l'état 1 et venant des divers états et en conséquence la somme des variances, soit :

$$\sum_{i=1}^{p} V^i\ a_{i1}(1 - a_{i1})$$

La matrice A ayant des termes a_{ij} voisins de zéro ($i \neq j$) ou de un ($i = j$) on voit facilement qu'en première approximation :

$$a_{i1}(1 - a_{i1}) \# a_{i1} \qquad \text{pour} \quad i \neq 1$$
$$a_{11}(1 - a_{11}) \# 1 - a_{11}$$

Ainsi la variance s'écrit :

$$V^1(1 - a_{11}) + \sum_{i=2}^{p} a_{i1} V^i$$

soit

$$V^1(1 - a_{11}) + (m - a_{11} V^1)$$

soit :

$$\sigma^2 = m + V^1 - 2 V^1 a_{11}$$

Si l'on admet que $a_{11} = 0,95$ alors $\sigma^2 \# m - 0,9\ V^1$

Pour un processus évoluant sur T année

$$\sigma^2 = m + V^1(1 - 2a_{11}^T) \# m + V^1[1 - 2(1 + T(a_{11} - 1))]$$
$$= m + V^1[2T(a_{11} - 1) - 1] \qquad (B)$$

Cette dernière formule n'est valable que si la matrice A est triangulaire.

Soit à l'aile de la formule (A) soit à l'aide de la formule (B) plus simple mais moins précise on montre que les fluctuations d'échantillonnage sont dix fois plus grandes que les fluctuations numériques entre les solutions proposées.

3 - VALIDITE DE LA METHODE

En dépit de la remarque précédente, il y avait lieu de se demander si les écarts constatés ne sont pas dûs à la méthode qui pourrait ne pas être appropriée au problème. C'est pourquoi, lors d'un passage, nous avons fait l'expérience suivante : nous avons remplacé les vecteurs $(V_{t_i + T_i})$ observés par les (W) calculés et correspondant à la meilleure matrice A, soient V_{v_i} ces vecteurs. Alors les différences entre les composantes des vecteurs prévus et les composantes des vecteurs observés V_{v_i} sont pour 1000 essais de l'ordre de 0,1 % au maximum.

La méthode convient donc au problème et met alors en cause les données. L'impossibilité rencontrée à faire baisser la distance S peut simplement venir de fluctuations d'échantillonnage dans les données.

Cette méthode peut paraitre lourde et peu satisfaisante pour l'esprit. Cependant la programmation en est très simple. Il est hors de doute qu'un tel calcul exceptionnel résolu par une méthode plus élaborée serait plus couteux. En dix minutes de 7040 une bonne solution est effectivement déterminée. Si l'on admet que l'application d'une autre méthode peut diviser ce temps par 10, cela représente un gain de l'ordre de la centaine de francs. Une telle somme ne permet pas de justifier une programmation plus complexe. Si l'on devait envisager une cinquantaine d'exploitation alors seulement le problème se poserait.

Vecteurs donnés

V_{t_1}	0,0123	0,0877	0,0395	0,1062	0,0901	0,3037	0,3605
V_{t_2}	0,0161	0,0907	0,0435	0,1193	0,0894	0,3118	0,3292
V_{t_3}	0,0251	0,0940	0,0476	0,1316	0,0927	0,3208	0,2882
V_{t_1+7}	0,0327	0,1080	0,0892	0,1608	0,1319	0,2425	0,2349
V_{t_2+6}	0,0351	0,1104	0,0891	0,1593	0,1317	0,2422	0,2321
V_{t_3+4}	0,0327	0,1056	0,0899	0,1609	0,1320	0,2426	0,2363

Vecteur de départ pour les prévisions

V_{62}	0,0421	0,1237	0,0918	0,1543	0,1301	0,2385	0,2194

Estimations

Après 689 essais on trouve la meilleure matrice $\hat{A}$ suivante. Les 1000 – 689 = 311 essais ultérieurs n'ont pas donné une matrice meilleure.

Matrice $\hat{A}$

1

0,00807 0,99193

0,00047 0,00158 0,99795

0,00287 0,00645 0,00228 0,98841

0,00217 0,00057 0,01540 0,00710 0,97476

0,00395 0,00839 0,01859 0,00271 0,03321 0,93316

0,00185 0,00084 0,00458 0,02362 0,00158 0,02430 0,94321

Vecteur prévu W_{t_1+7}

0,03374 0,10691 0,09771 0,15670 0,13654 0,22896 0,23943

Vecteur écart : V_{t_1+7} observé et W_{t_1+7} prévu par $\hat{A}$

0,00108 −0,00113 0,00852 −0,00410 0,00463 −0,01350 0,00450

Vecteur prévu W_{t_2+6}

0,03471 0,10781 0,09382 0,15941 0,13167 0,24078 0,23180

Vecteur écart : V_{t_2+6} observé et W_{t_2+6} prévu par $\hat{A}$

−0,00042 −0,00260 0,00473 0,00007 −0,00007 −0,00138 −0,00032

Vecteur prévu W_{t_3+4}

0,03760 0,10601 0,08202 0,15625 0,12360 0,26639 0,22812

Vecteur écart : V_{t_3+4} observé et W_{t_3+4} prévu par $\hat{A}$

0,00491 0,00042 −0,00786 −0,00465 −0,00839 0,02378 −0,00821

La prévision 70

Prévision 70 : V_{70} prévu à partir de V_{62}

0,06656 0,14100 0,14775 0,18533 0,15748 0,16444 0,13743

Moyenne des V_{70} pour les 10 matrices $\hat{A}$ rencontrées

0,06474 0,14129 0,14412 0,18424 0,16151 0,16611 0,13800

Bornes inférieures des V_{70} pour les 10 matrices $\hat{A}$

0,06223 0,13732 0,14202 0,18048 0,15748 0,16330 0,13348

Bornes supérieures des V_{70} pour les 10 matrices $\hat{A}$

0,06669 0,14518 0,14775 0,18558 0,16736 0,16892 0,14431

BIBLIOGRAPHIE

R. CASAMITJANA - Etude de la fonction matricielle P . Théorie et applications économiques. Thèse présentée en octobre 1966 à la Faculté des Sciences de Paris (113 pages).

A. BLANC, LAPIERRE et R. FORTET - Théorie des fonctions aléatoires Masson,1953,(693 pages).

A. KORGONOFF, L. BOSSET, J. L. GROBOILLOT, J. JOHNSON - Méthodes de calcul numérique Dunod, 1961, (375 pages).

F. DOUSSET et R. CASAMITJANA - Utilisation d'un modèle à processus stochastique pour la prévision de l'équipement des ménages en appareils de chauffage, congrès de la société d'économétrie, Varsovie 1966.