D.E.A SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION

UNIVERSITE LYON2 UNIVERSITE LYON3 ENSSIB

CIRCULATION DE L'INFORMATION: LE MODELE DE MORSE

Marie-Pierre ORIOL

Sous la direction de Thierry LAFOUGE

1993

D.E.A SCIENCES DE L'INFORMATION ET DE LA COMMUNICATION

UNIVERSITE LYON2 UNIVERSITE LYON3 ENSSIB

CIRCULATION DE L'INFORMATION: LE MODELE DE MORSE

Marie-Pierre ORIOL

Sous la direction de Thierry LAFOUGE

1993

Je tiens à exprimer toute ma reconnaissance à mon directeur de mémoire, Monsieur Thierry Lafouge, pour l'aide et le soutien qu'il a su m' accorder tout au long de l'année.

Je remercie également l'ensemble des enseignants du DEA pour leur compétence et leur disponibilité.

SOMMAIRE

I. Introduction

Tout bibliothécaire sait que ses rayons, accessibles ou non au public, sont encombrés par des ouvrages qui ne sont jamais demandés. Il peut présumer qu'ils ne répondent pas, ou plus, aux besoins des usagers, et qu'ils sont inutiles ou inutilisables. Ils le sont en effet le plus souvent, les ouvrages documentaires parce qu'ils contiennent des informations dépassées, les ouvrages de fiction parce qu'ils ne répondent plus aux impératifs de la mode ou de l'actualité et que la publicité et les mass-media ont cessé de les promouvoir.

Les accroissements des bibliothèques ne peuvent se développer de façon indéfinie. Ils sont soumis à des contraintes d'ordre commercial (disponibilités sur le marché de la librairie), d'ordre budgétaire, d'ordre intellectuel (vocation et niveau de la bibliothèque, qualité des utilisateurs). Ils trouvent aussi leurs limites dans la capacité de stockage des bâtiments.

"*Il est plus aisé d'amasser tous les livres que de les abréger et de les choisir*". Près de deux siècles plus tard, Alfred Sauvy faisait, non sans malice, écho à cette réflexion de Louis de Lacretelle :"*Egalement éprouvants sont les métiers de conservateurs et de révolutionnaires, car dans les deux cas, la difficulté est de savoir que détruire et que conserver*". La question n'a en effet guère progressé depuis le dix-huitième siècle, et la pratique bibliothèconomique n'y a pas encore apporté de réponse satisfaisante.

Pour gérer leurs stocks, les bibliothécaires qualifiés s'en remettent souvent à un empirisme raisonné qui peut se révéler efficace. Mais c'est parce que les évaluations jouent un rôle important dans la prise de décision, qu'il est apparu souhaitable au fil des ans d'élaborer des outils, appelés indicateurs, destinés à fonder les jugements sur des données observables, vérifiables et contrôlables. En particulier, les indicateurs bibliométriques prennent en compte les aspects quantitatifs des processus de production, de diffusion et de l'utilisation de l'information enregistrée. Ils permettent de dégager des tendances concernant la circulation des livres ou des périodiques, le comportement des lecteurs, le nombre de citations, etc...Par l'observation des fréquences d'événements appelées généralement distributions bibliométriques, de nombreux chercheurs en Science de l'information ont tenté, à l'aide de l'outil statistique, d'ajuster au mieux ces courbes par des modèles théoriques.

La méthode de Morse est une technique de micro-évaluation. Elle mesure une des activités les plus importantes des bibliothèques, le prêt des ouvrages, et donne des indications sur l'ensemble de la circulation des volumes par catégorie étudiée. Elle permet ainsi d'évaluer l'efficacité de la politique d'acquisition

Bien que le modèle de Morse ne présente pas de difficultés majeures si l'on se limite à la simple application de la méthode, son approche théorique est relativement complexe.

L'étude exposée dans ce mémoire a donc pour objectif l'analyse des fondements théoriques du modèle de Morse, l'explicitation de ses principaux résultats et de ses implications. On étudiera également les modifications ultérieures qui ont été apportées au modèle afin d'améliorer sa fiabilité.

II. Problématique de la circulation de l'information

. *Bibliométrie, scientométrie, infométrie*

La caractéristique de la bibliométrie est d'établir des études de publications sur des données quantitatives et non plus simplement subjectives. Ces données quantitatives sont calculées à partir de comptage statistiques de publications ou d'éléments extraits de ces publications.

On attribue à Pritchard [PRIT69] l'invention du terme **bibliométrie** qu'il a proposé en remplacement de son ancienne désignation, **bibliographie statistique,** qui pouvait prêter à confusion et être interprétée comme une bibliographie sur la statistique.Il a défini la bibliométrie comme étant

> *"l'application de méthodes mathématiques et statistiques aux livres et aux médias de communication".*

Cette définition de Pritchard ne donne aucune indication sur la finalité de la bibliométrie. A l'époque, son application s'insérait dans le domaine de la gestion des bibliothèques comme le montre la définition qu'en a donné Raising en 1962 [RAIS62] alors qu'elle était encore connue sous le nom de bibliographie statistique:

> *"l'assemblage et l'interprétation de statistiques relatives aux livres et aux périodiques...pour démontrer des mouvements historiques, pour déterminer l'utilisation par la recherche nationale et universelle des livres et des journaux, et pour s'assurer dans de nombreuses situations locales de l'utilisation des livres et des journaux".*

Depuis, la bibliométrie a fortement repoussé son application au-delà des frontières de la bibliothéconomie. Hawkins [HAWK77] plus récemment définit la bibliométrie comme "*les analyses quantitatives des caractéristiques bibliographiques d'un corps de littérature*". Mais par cette définition, une des activités de la bibliométrie n'est pas prise en compte : l'étude de la circulation des données.

Un autre terme, la **scientométrie**, est apparu parallèlement à celui de bibliométrie. Il est originaire d'un terme russe signifiant l'application de méthodes quantitatives pour l'histoire des sciences (Dobrov & Korennoi 1969 [DOBR69]).

Tandis que la bibliométrie aurait pour objet les livres ou les revues scientifiques et pour objectif les activités de communication de l'information, la scientométrie aurait pour objet les aspects quantitatifs de la création. diffusion et utilisation de l'information scientifique et technique et pour objectif la compréhension des mécanismes de la recherche comme activité sociale.

Donc la bibliométrie regrouperait les méthodes d'aide à la gestion des bibliothèques et la scientométrie rechercherait les lois qui régissent la science d'où son appellation "science de la science" par Price.

Un troisième terme l'**infométrie** serait le terme générique embrassant *bibliométrie* et *scientométrie*. Sa définition est bien plus large : l'infométrie est l'application des modèles et des méthodes mathématiques et statistiques de façon à dégager des lois relatives à l'information scientifique et technique.

Les différentes spécialités que l'on trouve en bibliométrie peuvent être découpées selon la liste donnée par H. Rostaing [ROST93]:

- **modélisation de distributions bibliométriques** (lois Bradford, Lotka et Zipf et des notions sur l'avantage du cumul)
- **indicateurs univariés** (mesures quantitatives basées sur des calculs de ratio)
- **indicateurs relationnels** (analyses statistiques descriptives des relations entre les éléments étudiés et qui donnent des indications plus qualitatives)
- **analyse bibliométrique des brevets** (applications des méthodes bibliométriques aux références brevets)
- **modélisation mathématique de la circulation des livres** (lois sur la diffusion et la communication des ouvrages)

Le dernier point fait l'objet de notre étude : par l'élaboration d'indicateurs de tendance, les modèles appliqués à la circulation des ouvrages dans une bibliothèque servent à l'analyse prévisionnelle de la demande et intéressent les gestionnaires des SID (Systèmes d'Information Documentaire). En effet, les bibliothèques ne peuvent souvent s'autoriser qu'une croissance limitée de leur fonds. Seule une politique active de désherbage des collections peut réussir à endiguer le déferlement de publications de ces dernières decennies. Le manque de place n'est pas la seule raison pour laquelle un bibliothécaire souhaite

"désherber" : laisser en rayon des livres peu ou pas réclamés peut déprécier une collection en la rendant inadéquate à la demande.

Bien entendu, comme le précise B. Richter [RICH92], la révision critique des collections n'a pas à être pratiquée dans toutes les bibliothèques. La Bibliothèque Nationale ne saurait l'envisager sans manquer à sa mission statutaire. Il en va de même pour les fonds régionaux et locaux des bibliothèques municipales qui assurent la collecte et la conservation des documents intéressant leur zone géographique d'activité. L'élimination apparaît, en revanche, indispensable dans les bibliothèques au service de la recherche théorique ou appliquée, dans les centres de documentation orientés vers la pratique, dans les bibliothèques d'enseignement, dans le fonds de vulgarisation des bibliothèques publiques.

Depuis les années 1960, un certain nombre de méthodes scientifiques de gestion des collections [BERT90] ont été proposées et testées, essentiellement dans les pays anglo-saxons dont elles sont souvent originaires. Des mathématiciens et spécialistes dont Morse, Burrell, Egghe, Rousseau, ont cherché à concevoir des modèles qui rendent compte à la fois de l'obsolescence qui affecte les collections et des phénomènes de regain de popularité (ou résurgence) qui peuvent toucher certains livres. De plus, ces modèles doivent pouvoir s'appliquer à tous les types de bibliothèques pendant que la valeur des paramètres utilisés diffère d'une biliothèque à l'autre. Les chercheurs ont montré que le critère le plus fiable était l'utilisation passée d'un document (fréquence des prêts, des consultations sur place; mesure du temps en rayon). Des approches privilégiant, soit le sujet de l'ouvrage, soit l'âge de l'information fournie (date d'édition de la publication originale, date d'acquisition de l'ouvrage) ont été rejetées car beaucoup moins fiables.

B. Obsolescence, mémoire et résurgence

C'est le triptyque-clé d'une analyse bibliométrique de la demande, celui qui inscrit la problématique d'un modèle à saisir un phénomène aussi divers et aléatoire que peut être la vie d'un ouvrage.

L'**obsolescence** marque le déclin de la demande relative à un ouvrage, à mesure que le temps passe depuis son édition (ou sa mise en circulation dans une bibliothèque donnée). C'est un phénomène général qui affecte la plupart des ouvrages -le cas des "classiques" pouvant être considéré comme marginal- .
R.Ducasse [DUCA78] précise que, selon la terminologie de Buckland, l'observation de l'obsolescence peut être effectuée à deux niveaux:

— le **niveau diachronique,** qui consiste à analyser sur une longue période la tendance de l'évolution de la demande relative à un ouvrage donné; dans la majorité des cas, cette évolution connaît une régression exponentielle négative dont le coefficient est plus ou moins élevé selon le poids de certains facteurs déterminants comme l'auteur, la langue, la discipline ou le genre.

— le **niveau synchronique**, qui consiste à analyser, à un moment donné, non plus l'évolution mais la structure de la demande selon l'âge des documents. Ajustés par des distributions théoriques, les résultats expérimentaux traduisent habituellement le fait que les ouvrages les plus anciens (dix ans ou plus) forment les effectifs relatifs les plus faibles.

La prise en compte par le gestionnaire d'un fonds documentaire du phénomène d'obsolescence est intéressante à plus d'un titre. Les paramètres de régression sont des indicateurs du vieillissement des collections et de l'oubli dans lequel tombent certains titres.De plus la mesure de l'obsolescence est essentielle à l'appréciation objective de l'état actuel de la circulation d'un ouvrage (ou d'une catégorie) par rapport à sa circulation passée et sa circulation attendue.

Malheureusement,en s'appuyant sur des méthodes trop rigides qui accordent à toutes les valeurs passées le même poids, la mesure de l'obsolescence est un piètre outil de prévision à court terme ; on dira qu'elle manque de *mémoire immédiate.*

Ce concept de mémoire immédiate est très important en matière de prévision. En effet,parmi les nombreuses procédures d'analyse de données nécessitant un

ensemble de valeurs passées de grande dimension, rares sont celles qui parviennent, ainsi le lissage exponentiel, à accorder aux valeurs récentes une prépondérance suffisante pour déterminer avec une fiabilité acceptable la valeur que prendra la variable dans un futur immédiat. On pourrait dire en se plaçant au temps t, qu'elles donnent généralement à la valeur initiale obtenue au temps t_0 la même importance qu'à la valeur obtenue au temps t-1.

Or, en toute hypothèse, il est probable que la corrélation entre les valeurs prises au temps t et au temps t-1 soit plus forte que celle qui pourrait exister entre t_0 et t, surtout si l'intervalle $[t_0, t]$ est grand.

Il s'agit donc désormais de rechercher une méthode d'analyse qui intègre la valeur que prend la variable au temps t-1 avec tout son poids. Cette méthode, compte tenu de ce qui précède, serait aussi la seule à permettre la mise en évidence, en temps importun, d'un phénomène de résurgence. La **résurgence** correspond au cas où la valeur de la variable rompt de manière indiscutable avec ses valeurs précédentes -qu'elles soient appréciées en termes de tendance ou de moyenne-. Un tel phénomène étant somme toute assez fréquent, un ouvrage pouvant pour des raisons très diverse être soudainement très demandé ou au contraire délaissé, nous verrons qu'un processus aléatoire, dit de Markov, est à même d'en rendre compte de façon tout à fait satisfaisante.

III. Le modèle markovien de Morse

Circulation des livres et processus de Markov

1.Observation empirique de fréquences de circulation

Présenté pour la première fois dans *Library Effectiveness* [MORS68], ouvrage fondamental paru en 1968, le modèle markovien de Morse est généralement considéré par les spécialistes anglo-saxons comme le point de départ des applications de la recherche opérationnelle à la gestion des systèmes d'information de type bibliothèque. Parmi les études présentées jusqu'alors, le modèle de Morse s'imposait comme le seul modèle probabiliste approprié à la circulation des livres.

Le modèle markovien de Morse est basé sur l'observation expérimentale, pendant une longue période, de la circulation d'une série d'ouvrages tirés au hasard, opération qui met en avant les phénomènes d'obsolescence et de résurgence dont nous avons parlé précédemment.

Morse a relevé dans un tableau (tableau 3.1) les données relatives à la circulation annuelle de quatre ouvrages A, B, C, D, tel que les chiffres de chaque ligne représentent le nombre d'emprunts annuel sur 25 années successives

tableau 3.1: Chroniques de la circulation annuelle de 4 ouvrages A,B,C,D appartenant à la bibliothèque scientifique du MIT [MORS68, p.86]

Ouvrage A	8 5 3 3 1	2 0 2 2 2	1 1 0 1 1	3 2 3 0 1	2 0 1 1 0
Ouvrage B	4 1 0 1 1	1 2 4 0 3	2 2 2 1 0	0 1 0 0 2	1 1 0 0 0
Ouvrage C	2 1 0 0 0	0 1 3 4 3	3 4 5 3 0	0 3 1 2 0	1 0 1 0 0
Ouvrage D	1 0 1 0 0	1 0 0 0 0	1 1 1 0 1	0 1 0 0 0	0 0 0 0 0

Morse analyse sommairement ces chroniques en notant qu'en dépit d'une tendance générale à la baisse, elles évoluent de manière fort différente; la représentation graphique du tableau 3.1 (voir figures 3.1et 3.2) permet de mettre en évidence ces disparités :

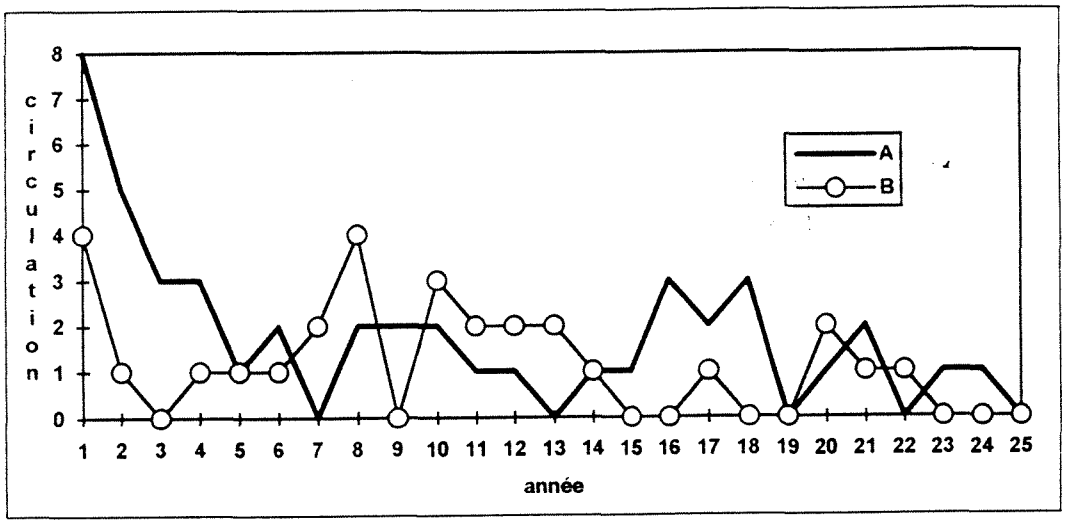


figure 3.1: courbes de circulation des ouvrages A et B

Les chroniques des ouvrages A et B décroissent au fur et à mesure que le temps s'écoule, ce qui est le cas de la plupart des ouvrages. Bien que la circulation annuelle présente d'indéniables fluctuations (on peut observer une alternance de pics d'obsolescence et de résurgence), l'aspect général est celui d'une tendance décroissante.

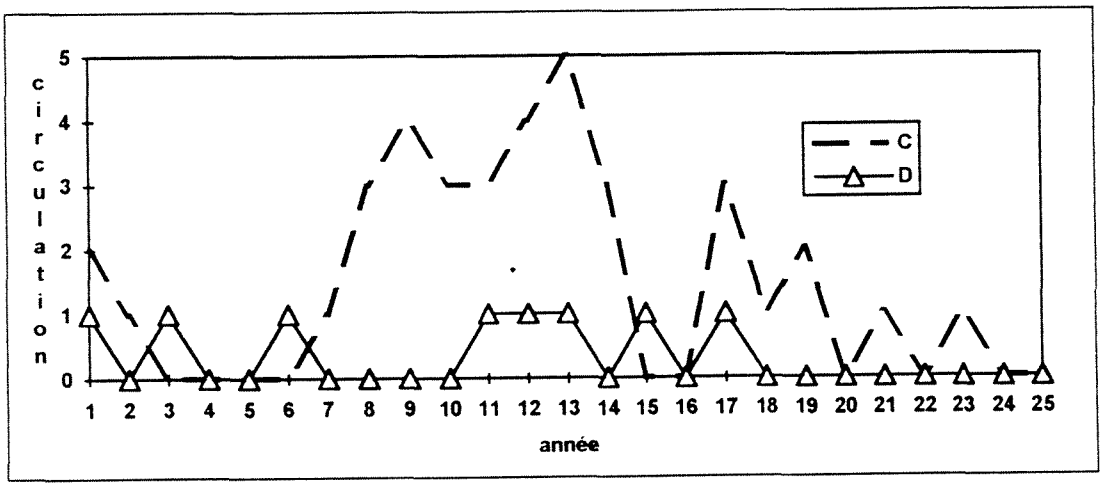


figure 3.2: courbes de circulation des ouvrages C et D

Le cas de l'ouvrage C est plus complexe et ne se rencontre qu'occasionnellement. Le livre connaît un regain d'intérêt après 8 ans passés en rayon, ainsi que cela arrive parfois à des monographies citées dans tel ou tel programme universitaire.Le taux de circulation décroît ensuite de façon similaire aux ouvrages A et B.

Quant à la chronique D, elle voit succéder à un démarrage très faible une période de 8 années durant laquelle l'ouvrage est à nouveau sollicité -mais reste cependant peu réclamé-, pour rechuter ensuite de façon apparemment inexorable.

C'est dans la mesure où de telles distributions seront ajustées de façon satisfaisante, tant du point de vue des données observées que de leur structure, que l'on admettra la validité du modèle markovien.

2. Hypothèses du modèle

Le modèle de Morse repose sur trois hypothèses fondamentales:

- On suppose que **la circulation des ouvrages est un processus aléatoire**. Autrement dit, l'emprunt de livres dans une bibliothèque est le résultat d'un grand nombre d'occurrences aléatoires. Le comportement de tout usager, ou le taux d'utilisation d'un livre donné ne peuvent être prédits avec certitude. Rien ne permet d'affirmer que tel livre sera emprunté sept fois l'année suivante, tandis que tel autre ne sera pas emprunté du tout. Néanmoins, il est possible de prédire le comportement *moyen* d'une classe[1] de livres, la précision de cette prédiction étant proportionnelle au nombre d'ouvrages de la classe considérée

- **En moyenne, la circulation des ouvrages décroît exponentiellement avec le temps**. Cette hypothèse a été tirée des études de Trueswell dans *College and Research Libraries* [True64] et de Fussler et Simon [Fuss61].

- **Il existe une correlation entre l'utilisation d'un livre pendant une période donnée et son utilisation pendant la période suivante**. En d'autres termes, bien qu'en moyenne la circulation d'un livre décroisse exponentiellement d'année en année, il y a beaucoup d'exceptions : des livres restés en rayon depuis plusieurs années peuvent devenir brusquement populaire. Dès qu'un livre devient populaire, sa circulation future se comporte dès lors comme s'il l'avait toujours été. De même , si un livre connaît une baisse soudaine de popularité, sa circulation future sera la même que s'il avait toujours été dédaigné.

Certains modèles appliqués aux opérations de bibiliothèque s'avèrent, malgré leur fondement théorique, inutilement compliqués et trop orientés sur les mathématiques, ce qui les rend difficiles d'accès aux bibliothécaires. C'est pourquoi Morse indique dans son article du Library Quarterly [MORS72] qu'il préfère utiliser un modèle nécessitant peu de données, mais pouvant conduire à

---

[1] on désigne par "classe" ou "catégorie" un ensemble d'ouvrages possédant une caractéristique commune qui peut être par exemple , la matière (chimie, physique, informatique ...), l'indice de classification Dewey , la date de mise en circulation...

une erreur de 25%, plutôt qu'un modèle plus fiable mais nécessitant des années-homme de travail pour son implantation:

"Puisque notre étude porte sur une grande quantité de livres, notre position doit plutôt être celle d'une compagnie d'assurance qui peut perdre sur quelques cas mais doit gagner sur l'ensemble[2]."

3.Flux poissonien d'événements

Puisque par hypothèse la demande d'ouvrage est un processus aléatoire, on peut l'approcher par une loi de Poisson. En effet, on dit qu'un processus temporel est un processus de Poisson s'il représente l'apparition d'événements aléatoires E_1, E_2, ...,E_n, etc., satisfaisant aux 3 conditions suivantes[3]:

1. *La loi du nombre d'événements N arrivant dans l'intervalle $\{t_0; t_0+T\}$ ne dépend que de T. Si T=1 on notera λ l'espérance-dite "cadence"- de cette loi.*
2. *Les temps d'attente E_1, E_2,...,E_n sont des variables indépendantes (processus sans mémoire).*
3. *Deux événements ne peuvent arriver simultanément.*

figure 3.3

Dans le cas du processus de demande d'ouvrage, un événement est représenté par l'emprunt d'un livre. Sachant qu'un même livre ne peut faire l'objet que d'un seul prêt à la fois et que les durées (temps d'attente) séparant deux prêts consécutifs ne sont pas dépendantes, les conditions 2 et 3 sont remplies. Quant à la première condition il est évident que la loi de demande d'ouvrage va dépendre de la longueur T de l'intervalle de temps considéré et non de l'instant initial t_0 .

---

[2]Philip M. Morse [MORS72,p.20]

[3]pour un développement théorique voir le chapitre V "Notions élémentaires sur les processus aléatoires"dans l'ouvrage *Probabilités, Analyse des données et Statistique* de Saporta [SAPO90, p.109]]

La demande d'ouvrage peut donc s'approcher par un processus poissonien; il s'ensuit que **le nombre d'événements se produisant pendant une période de durée T fixée suit une loi de Poisson de paramètre λT** :

Donc la probabilité d'avoir n événements (ou n demandes)pendant l'intervalle T est:

$$P_n(T)=(\lambda T)^n/n!\ \exp(-\lambda T) \qquad (3.1)$$

où λT est le nombre moyen d'événements réalisés dans cet intervalle de temps Quant à l'intervalle de temps D qui sépare deux demandes consécutives E_i, E_{i+1}dans un processus poissonien, la probabilité qu'il soit compris entre t et t+dt est égale au produit de la probabilité qu'il ne survienne aucune demande jusqu'à t, par la probabilité qu'il s'en produise une durant dt. D'où

$$P(t<D<t+dt)=P_0(t).P_1(dt)$$

$$=\exp(-\lambda t).(\lambda dt)\exp(\lambda dt)$$

Si dt est infiniment petit, $\exp(-\lambda t) \sim 1-\lambda t$, et $(\lambda dt)^2 \sim 0$,donc

$$P(t<D<t+dt)=\lambda\exp(-\lambda t)dt \qquad (3.2)$$

expression d'une loi exponentielle de moyenne $1/\lambda$ (temps moyen séparant deux demandes consécutives).

4.Processus de Markov

Afin de prendre en compte la corrélation qui existe entre les circulations de deux années consécutives, on doit insérer une mémoire dans ce processus aléatoire. Le **processus de Markov** est le plus simple processus stochastique qui rende compte d'une telle dépendance temporelle.

Dans un **processus markovien**, l'état d'un système à une période donnée n'est déterminé que par son état à la période précédente . L'information apportée par les états antérieurs est entièrement contenue dans la connaissance de l'état le plus récent.(☞Annexe A)

Un tel processus est basé sur l'estimation de **la matrice $[T_{mn}]$ des probabilités de transition**.Cette matrice a pour terme général la probabilité conditionnelle

T_{mn}= "Probabilité qu'un ouvrage circule n fois pendant l'année t+1 sachant qu'il a circulé m fois durant l'année t"

Morse fait l'hypothèse que la distribution des ouvrages empruntés sur deux années s'inscrit dans un **processus de Markov homogène d'ordre 1.** Plus précisément. il suppose que T_{mn} prend la forme suivante,

$$T_{mn} = \frac{(\alpha+\beta m)^n}{n!} e^{-(\alpha+\beta m)} \qquad (3.3)$$

probabilité conditionnelle d'une loi de Poisson de moyenne $\alpha+\beta m$ et qui satisfait la condition de normalisation :

$$\sum_{n=0}^{\infty} T_{mn}=1 \qquad (3.4)$$

On peut utiliser l'équation (3.3) pour calculer la quantité N(m), qui est le nombre moyen de circulations durant l'année t+1, sachant qu'il y a eu m circulations durant l'année t. Nous obtenons, en utilisant l'équation (3.3)

$$N(m) = \sum_{n=0}^{\infty} nT_{mn}$$

$$= \alpha+\beta m \qquad (3.5)$$

Cela signifie que la circulation moyenne d'une classe de livres durant l'année t+1 dépend uniquement de sa circulation durant l'année t précédente. La circulation moyenne de la (t+1)-ème année, N(m), ne dépend explicitement ni de l'âge de l'ouvrage, ni de sa circulation durant les années précédant la t-ième année.

Les paramètres α et β sont à déterminer[4] pour chaque classe de livres étudiée :

- α mesure la valeur de la circulation moyenne éventuellement atteinte par les ouvrages les plus anciens de la classe considérée.
- β mesure la diminution de "popularité" d'un livre d'une année sur l'autre

Généralement β reste constant durant toute la durée de vie d'un ouvrage tandis que α diminue légèrement avec le temps (voir tableau 3.2).Le paramètre β peut être très inférieur à l'unité et se situe habituellement entre 0,2 et 0,8. Le paramètre α peut être supérieur à l'unité mais est généralement compris entre 0,3 et 0,7

Morse notait en 1972 [MORS72] que "les données recueillies jusqu'à présent montrent que β reste en gros constant pendant 10 à 20 ans tandis que α tombe aux 2/3 de sa valeur après 10 à 20 ans de vie du livre à la bibliothèque".Nous verrons en fait au chapitre V que le paramètre α décroît linéairement avec le temps quand il est calculé sur de très grands échantillons de données. Mais comme la majorité des études réalisées jusqu'à présent n'ont porté que sur des ensembles de données restreints (faute de moyens de calcul puissants) on peut admettre l'affirmation de Morse ainsi que les résultats qui vont suivre.Il était d'ailleurs dans l'objectif de Morse de limiter au maximum le recueil des données.

Tableau 3.2: paramètres de Markov pour différentes classes de livres [MORS68, p.107]

Class of Book	β All Years	α First 4 Years	α About 8th Year	α About 12th Year
Biology	0.45	0.4	0.3	0.2
Chemistry	0.5	0.6	0.5	0.3
Engineering	0.7	0.3	0.2	0.1
Geology	0.2	0.4	0.2	0.1
Mathematics	0.6	0.3	0.25	0.2
Metallurgy	0.7	0.3	0.25	0.2
Physics	0.6	0.5	0.4	0.3
All Classes	0.5	0.4	0.25	0.15

[4]Les diverses méthodes de calcul des paramètres α et β sont exposées au chapitreV

5.La relation linéaire N(m)=α+βm

Nous allons maintenant expliciter le processus expérimental utilisé pour obtenir la relation N(m)=α+βm (3.5). Il est d'abord nécessaire de définir les quantités que l'on peut tirer de l'observation des circulations d'ouvrages sur deux années successives (cf. tableau 3.3 page suivante). Afin d'illustrer les relations existant entre ces diverses quantités, quelques données expérimentales et théoriques relevées dans la thèse de Chen [CHEN76] sont présentées dans le tableau 3.4, qui représente la circulation de la classe WM (psychiatrie) des livres qui ont été retournés à la bibliothèque Countway de l'Université d'Harvard pendant le mois de Janvier 1973. Les données sont relatives aux circulations de M=560 ouvrages, observées durant les années 1972, 1973.

En ce qui concerne le recueil des données Morse a spécifié que l'on peut obtenir une précision suffisante en examinant quelques centaines de fiches de prêt d'une classe de livres donnée. Les fiches examinées doivent être sélectionnées à partir d'un échantillonage au hasard : par exemple tous les deux livres, ou tous les trente livres si l'on veut couvrir tous les rayonnages occupés par la classe considérée et si l'on veut que l'échantillon comporte plusieurs centaines de fiches. Il n'est pas nécessaire d'étudier toutes les classes le même mois.

Tableau 3.4 : Valeurs de M(m), N_{mn}, et N(m) pour différentes valeurs de m et n [CHEN76, p.14]

m	$M(m)$	N_{mn}													$N(m)$	Theo-retical
		n=0	1	2	3	4	5	6	7	8	9	10	11	12		
0	91	22	26	17	13	9	2	0	1	0	0	1	0	0	1.78	*1.58*
		18	*29*	*23*	*12*	*4*	*4*	*0*	*0*	*0*	*0*	*0*	*0*	*0*		
1	109	22	33	22	16	6	7	2	0	1	0	0	0	0	1.87	*1.95*
		15	*30*	*29*	*19*	*9*	*6*	*3*	*1*	*0*	*0*	*0*	*0*	*0*		
2	116	15	32	22	22	16	3	4	1	1	0	0	0	0	2.24	*2.31*
		11	*26*	*30*	*23*	*13*	*6*	*2*	*0*	*0*	*0*	*0*	*0*	*0*		
3	83	8	18	19	20	11	2	3	1	1	0	0	0	0	2.45	*2.68*
		5	*15*	*20*	*18*	*12*	*6*	*2*	*1*	*0*	*0*	*0*	*0*	*0*		
4	73	5	7	18	15	12	10	1	2	2	1	0	0	0	3.16	*3.05*
		3	*10*	*16*	*16*	*12*	*7*	*3*	*1*	*0*	*0*	*0*	*0*	*0*		
5	32	2	3	4	8	6	2	1	1	2	3	0	0	0	3.91	*3.41*
		1	*3*	*6*	*6*	*5*	*4*	*2*	*1*	*0*	*0*	*0*	*0*	*0*		
6	22	1	2	5	1	9	1	0	2	0	0	1	0	0	3.64	*3.78*
		0	*1*	*3*	*4*	*4*	*3*	*2*	*1*	*0*	*0*	*0*	*0*	*0*		
7	9	0	3	2	0	1	2	0	0	1	0	0	0	0	3.22	*4.14*
		0	*0*	*1*	*1*	*1*	*1*	*1*	*0*	*0*	*0*	*0*	*0*	*0*		
8	14	0	3	1	0	2	2	2	2	1	0	0	0	1	4.93	*4.51*
		0	*0*	*1*	*2*	*2*	*2*	*1*	*1*	*0*	*0*	*0*	*0*	*0*		
9-13	11	...														
M = 560																

tableau 3.3 : liste des effectifs expérimentaux et théoriques

Notation	définition	Effectif	
		expérimental	*théorique*
N_{mn}	Nombre d'ouvrages ayant circulé m fois durant l'année t, et n fois durant l'année t+1	N°_{mn}	$N_{mn}=M^{\circ}(m)\,T_{mn}$
X_n	Nombre d'ouvrages ayant circulé n fois pendant l'année t+1	$X^{\circ}_n=\sum_{m\geq 0} N^{\circ}_{mn}$	$X_n=\sum_{m\geq 0} N_{mn}$
M(m)	Nombre d'ouvrages ayant circulé m fois pendant l'année t	$M^{\circ}(m)=\sum_{n\geq 0} N^{\circ}_{mn}$	$M(m)=\sum_{n\geq 0} N_{mn}$
M	Nombre total d'ouvrages d'une classe	$M=\sum_{m\geq 0} M^{\circ}(m)$	
N(m)	Nombre moyen de circulations durant l'année t+1 sachant que l'ouvrage a circulé m fois durant l'année t	$N^{\circ}(m)=[1/M^{\circ}(m)]\sum_{n\geq 0} nN^{\circ}_{mn}$	$N(m)=\sum_{n\geq 0} nT_{mn}$

Le tableau 3.4 donne à la fois les résultats expérimentaux et théoriques de N_{mn} et N(m) (les données théoriques sont en italiques).

Si l'on trace sur un graphe les valeurs de N°(m) (valeurs expérimentales) en fonction de m (cf. figure 3.4) pour le couple d'années 1971-1972, on s'aperçoit que l'on peut tracer une droite qui ajuste au mieux le nuage de points. Ceci prouve graphiquement que la relation existant entre N(m) et m est une fonction linéaire, qui peut donc se formuler par l'équation (3.5)

$$N(m)=\alpha+\beta m$$

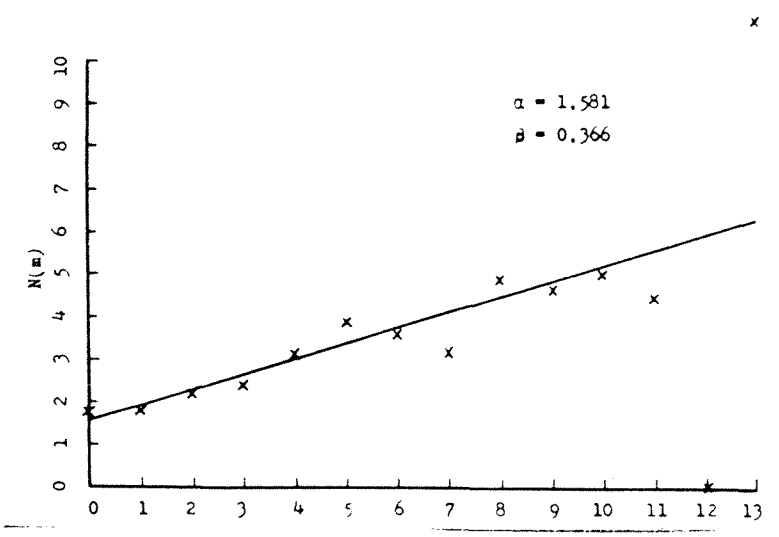


Figure 3.4 : Circulation moyenne N(m) pour l'année t+1 en fonction de la circulation m de l'année t précédente (données du tableau 3.4)

Dans l'exemple donné ci-dessus, les valeurs de α et β trouvées par la méthode des moindres carrés pondérés[5] étaient respectivement 1,581 et 0,366. Une fois connues les valeurs de α et β, et étant donné m, on peut aisément calculer les valeurs théoriques de N(m) en utilisant l'équation (3.5) : $N(m)=\alpha+\beta m$. Une comparaison des valeurs expérimentales et théoriques de N(m) du tableau 3.4 semble montrer que l'équation (3.5) permet de faire une bonne approximation.

[5]voir chap. V, méthodes de calcul

B. Prévision de la circulation future

1. Circulation moyenne d'une collection

Une fois calculées les valeurs de α et β de l'évolution des prêts en fonction du temps pour une catégorie déterminée de livres dans une bibliothèque donnée, on peut en déduire un grand nombre de renseignements sur les prêts des livres de cette catégorie . Le modèle de Markov nous permet de prévoir la valeur de la circulation moyenne, R(t+1), pour la même collection durant l'année à venir (t+1), en fonction de la valeur observée R(t) pour l'année écoulée et de celles des paramètres α et β qui restent pratiquement constantes durant plusieurs années (aussi longtemps que la composition ou l'emplacement de la collection ne sont pas radicalement modifiés).

Dans l'article *Measures of Library Effectiveness* [MORS72], Morse donne le résultat suivant s'appliquant à une collection déterminée de livres (sans adjonction de livres nouveaux):

> La circulation moyenne d'une collection durant l'année t+1 est liée à la circulation moyenne de l'année précédente t par la formule :
>
> $$\mathbf{R(t+1)=\alpha+\beta R(t)} \qquad (3.6)$$

On peut proposer la démonstration suivante de ce résultat:

■ Si un livre d'une classe donnée a circulé m fois pendant une année t, alors sa circulation moyenne théorique durant l'année t+1 sera d'après l'équation (3.5)

$$N(m)=\alpha+\beta m$$

Ce qui peut encore s'écrire, en utilisant les notations du tableau 3.3

$$[1/M°(m)] \sum_{n\geq 0} n\, N_{mn} = \alpha+\beta m \qquad \forall\ m > 0$$

avec $M°(m)=\sum_{n\geq 0} N°_{mn}$

En multipliant l'équation précédente par l'expression de M°(m), on obtient l'expression de la circulation totale pour l'année t+1 des livres ayant circulé m fois pendant l'année t:

$$\sum_{n\geq 0} n\, N_{mn} = \alpha \sum_{n\geq 0} N°_{mn} + \beta m \sum_{n\geq 0} N°_{mn}$$

En effectuant ensuite une sommation d'indice m, on obtiendra la circulation totale estimée d' une classe de livres pour l'année t+1

$$\sum_{m\geq 0} \left(\sum_{n\geq 0} n\, N_{mn}\right) = \alpha \sum_{m\geq 0} \left(\sum_{n\geq 0} N°_{mn}\right) + \beta \sum_{m\geq 0} \left(\sum_{n\geq 0} m N°_{mn}\right)$$

Pour avoir la circulation moyenne de cette classe de livres, on doit diviser la circulation totale par le nombre d'ouvrages de la classe, c'est-à-dire par $M=\sum_{m\geq 0} M°(m)$

$$= \sum_{m\geq 0} \left(\sum_{n\geq 0} N°_{mn}\right)$$

On arrive finalement à la relation

$$(1/M) \sum_{m\geq 0} \left(\sum_{n\geq 0} n\, N_{mn}\right) = \alpha + (1/M).\beta \sum_{m\geq 0} \left(\sum_{n\geq 0} m N°_{mn}\right)$$

ce qui est bien l'expression de la relation (3.6)

$$R(t+1)=\alpha+\beta R(t)$$ ■

Notons que si la relation (3.5) implique la relation (3.6), la réciproque (3.6)⇒(3.5), par contre, n'est pas vraie.

Si le paramètre β est de peu inférieur à l'unité, la circulation moyenne durant l'année t+1 sera inférieure de très peu à celle de l'année passée. Mais si β est très inférieur à l'unité, la circulation moyenne chutera considérablement au cours des premières années de vie du livres.

Si les paramètres α et β sont supposés constants au cours des années et si l'on fixe une origine des temps à t=1, alors d'après la relation (3.6) la circulation moyenne d'une classe de livres à la date t=2 sera

$$R(2)=\alpha+\beta R(1)$$

Durant la troisième année, cette même classe aura comme circulation moyenne,toujours d'après l'équation (3.6)

$$R(3)=\alpha+\beta R(2)$$

$$=\alpha+\beta[\alpha+\beta R(1)]$$

$$=\alpha(1+\beta)+\beta^2 R(1)$$

En raisonnant par itérations successives, on obtiendra pour la (t+1)ème année la circulation moyenne

$$R(t+1) = \alpha(1+\beta+\beta^2+\ldots+\beta^{t-1}) + \beta^t R(1)$$

$$R(t+1)= \alpha \frac{(1-\beta^t)}{1-\beta} + \beta^t R(1) \qquad (3.7)$$

Comme $\beta^t= \exp[-t \ln(\beta^{-1})]$ si $\beta>0$, on peut encore écrire (3.7) sous la forme

$$R(t+1)= \alpha/(1-\beta) + [R(1) - \alpha/(1-\beta)]\, \beta^t$$

$$\mathbf{R(t+1)= \alpha/(1-\beta) + [R(1) - \alpha/(1-\beta)].\exp[-t \ln(\beta^{-1})]} \qquad (3.8)$$

Ces relations nous permettent donc d'estimer la circulation moyenne d'une classe d'ouvrages à l'instant t+1 connaissant la circulation moyenne à l'instant 1.

La fonction $t \rightarrow R(t+1)$ décroît exponentiellement si $\mathbf{0<\beta<1}$ et si $R(1)-\alpha/(1-\beta)>0$, autrement dit si

$$R(1)>\alpha+\beta R(1)$$

soit **R(1)>R(2)**

Sous ces conditions, l'équation (3.8) montre que la circulation des ouvrages décroît exponentiellement avec le temps comme le formulait la deuxième hypothèse fondamentale du modèle.

Bien que la circulation moyenne d'une série donnée de livres décroisse avec le temps il s'avère cependant que dans beaucoup de bibliothèques elle reste passablement constante parce que les volumes anciens sont en permanence remplacés par des livres nouveaux davantage demandés. Par conséquent, on peut admettre, comme le suggère Morse, que les chiffres R, la circulation moyenne de l'ensemble des volumes d'une classe, et C, la fraction active[6] de cette classe, sont plus ou moins indépendants du temps. Ce qui change, c'est le total des livres de la classe.

La circulation moyenne R, et la fraction active, C, des nouveaux livres peuvent être obtenues des données concernant les livres mis en rayon depuis moins de deux ans. Ces quantités permettent de mesurer l'efficacité de la politique d'achat. C'est cette distinction qu'a faite Simon Cane lors de la mise en place du modèle de Morse à la bibliothèque de lecture publique d'Autun [CANE87]:

"*La mise en oeuvre de cette équation* [équation 3.6] s*uppose que l'on distingue le comportement des nouveautés, les volumes acquis depuis moins de deux ans, et celui des volumes disponibles depuis au moins deux ans, les ouvrages "anciens"....En ce qui concerne les nouveautés on effectue les mesures C et R qui constituent des indicateurs de la politique d'acquisition. Sur les volumes anciens on effectue les mêmes mesures afin de connaître le pourcentage de volumes actifs et le taux de rotation pour l'ensemble des volumes de chaque catégorie et de pouvoir comparer ces indicateurs à ceux des nouveautés.Les données sur l'ensemble des volumes permettent d'effectuer des prévisions sur le comportement de la classe étudiée....Le calcul de* α *et* β *s'effectue uniquement à partir des mesures du comportement des anciens volumes."*

2.Influence du temps sur la circulation

Si α et β demeurent constants durant la durée de vie des ouvrages, alors l'équation (3.7) peut prédire la circulation moyenne de cette classe de livres pour l'année t ou t+1. Néanmoins, les résultats expérimentaux montrent clairement que α diminue lentement avec le temps. Après T années, la valeur

---

[6]La fraction active représente la proportion de livres qui ont circulé au moins une fois pendant l'année considérée.On trouvera son expression au chap VI "Les distributions géométriques" (§ B)

de α se changera en α', et la circulation moyenne pour l'année t+1 (t>T) sera

$$\mathbf{R(t+1) = 1/(1-\beta)\,[\alpha' + (\alpha-\alpha')\beta^{t-T+1} - \alpha\beta^{t}] + \beta^{t}\,R(1)} \qquad (3.9)$$

■ Démonstration : si pendant T années la valeur de α ne varie pas, la circulation moyenne pour l'année T d'une classe de livres sera d'après la relation (3.7)

$$R(T) = \alpha\,\frac{1-\beta^{T-1}}{1-\beta} + \beta^{T-1}\,R(1)$$

A partir de l'année T+1 la valeur de α se change en α'. Donc la circulation moyenne de l'année T+1 peut s'écrire en fonction de celle de l'année T

$$R(T+1) = \alpha'+\beta R(T) \qquad \text{d'après (3.6)}$$

$$= \alpha'+\beta\left[\alpha\,\frac{1-\beta^{T-1}}{1-\beta} + \beta^{T-1}\,R(1)\right]$$

De même la circulation moyenne de l'année T+2 s'écrira

$$R(T+2) = \alpha'+\beta R(T+1)$$

$$= \alpha'+\beta\left(\alpha'+\beta\left[\alpha/(1-\beta)\,(1-\beta^{T-1}) + \beta^{T-1}\,R(1)\right]\right)$$

$$= (1+\beta)\alpha' + \alpha\beta^{2}\cdot(1-\beta^{T-1})/(1-\beta) + \beta^{T+1}\,R(1)$$

Et l'on obtient comme circulation moyenne pour l'année T+3

$$R(T+3) = \alpha'+\beta R(T+2)$$

$$= \alpha'+\beta(1+\beta)\alpha' + \alpha\beta^{3}\cdot(1-\beta^{T-1})/(1-\beta) + \beta^{T+2}\,R(1)$$

$$= \alpha'(1+\beta+\beta^{2})+\alpha\beta^{3}/(1-\beta) - \alpha\beta^{3+T-1}/(1-\beta)+\beta^{T+2}\,R(1)$$

$$= \alpha'(1-\beta^{3})/(1-\beta)+\alpha\beta^{3}/(1-\beta)-\alpha\beta^{3+T-1}/(1-\beta)+\beta^{T+2}R(1)$$

$$= \alpha'/(1-\beta)-\alpha'\beta^{3}/(1-\beta)+\alpha\beta^{3}/(1-\beta)-\alpha\beta^{3+T-1}/(1-\beta)+\beta^{T+2}R(1)$$

$$R(T+3) = \alpha'/(1-\beta)+(\alpha-\alpha')\beta^3/(1-\beta)-\alpha\beta^{3+T-1}/(1-\beta)+\beta^{T+2}R(1)$$

On en déduit que la circulation moyenne pour l'année T+n (n>1) s'écrit

$$R(T+n) = \alpha'/(1-\beta)+(\alpha-\alpha')\beta^n/(1-\beta)-\alpha\beta^{n+T-1}/(1-\beta)+\beta^{n+T-1}R(1)$$

Il suffit ensuite de poser T+n=t+1 , de remplacer dans la relation précédente, n par t-T+1, et n+T-1 par t, pour obtenir la relation (3.9)

$$R(t+1) = 1/(1-\beta)\,[\alpha' + (\alpha-\alpha')\beta^{t-T+1}-\alpha\beta^t] + \beta^t R(1) \qquad \blacksquare$$

⇒<u>valeur asymptotique</u>: Les formules (3.8) et (3.9) montrent que la circulation moyenne d'une collection tend, lorsque $t\to\infty$, vers une valeur constante $\alpha/(1-\beta)$,ou $\alpha'/(1-\beta)$, (si $\beta<1$).Le paramètre β mesure la rapidité avec laquelle le taux de circulation approche $\alpha/(1-\beta)$,ou $\alpha'/(1-\beta)$.

Morse a relevé en 1962 dans la collection de la bibliothèque du MIT les circulations relatives à diverses classes s de livres, d'effectif N_S. Les circulations d'une centaine de livres de chaque classe ont été analysées pour obtenir les valeurs approchées α_S et β_S de chaque classe de livres (voir tableau 3.5) En multipliant chaque α_S par N_S, en sommant et en divisant par N, le nombre total de livres de la collection, on a obtenu les valeurs moyennes de α et β[7] (α=0.40 et β = 0.49) pour toute la collection

Tableau 3.5 : Paramètres de circulation pour différentes classes de livres [MORS68,p.95]

Class s	N_s	α_s	β_s	$N_s\alpha_s$	$N_s\beta_s$
General Science and Engineering	1400	0.30	0.70	420	980
Mathematics	4900	0.30	0.60	1470	2940
Physics	3700	0.50	0.60	1850	2220
Chemistry	2800	0.60	0.50	1680	1400
Geology	6700	0.45	0.20	3020	1340
Biology	5000	0.40	0.45	2000	2250
Metallurgy, Food Technology	3500	0.20	0.70	700	2450
M =	28000			11140	13580

[7]Les paramètres de markov d'une collection peuvent se calculer à l'aide des paramètres des diverses classes de la collection, grâce aux propriétés baycentriques (voir chap. V §A,2)

Nous pouvons tirer de ce tableau les observations suivantes: si l'on suppose que α'_S (valeur de α_S après T années) est proportionnelle à α_S, on peut constater que les livres de géologie perdent de leur popularité au bout de deux ans (car β_S est petit), et tendront rapidment à être empruntés en moyenne une fois tous les deux ans ($\alpha \cong 0,45$; $1-\beta \cong 0,8$; $0,45/0,8 \cong 0,56$) .

Les livres de mathématiques, par contre, ne perdent que lentement leur popularité initiale ($\beta = 0,6$), mais seront finalement empruntés moins d'une fois par an, puisque $\alpha = 0,3$ et $\alpha/(1-\beta) \cong 0,7$.

Les livres de chimie et de physique demeurent populaires relativement longtemps ($\beta \geq 0,5$) et leur taux de circulation tendra à être plutôt élevé puisque $\alpha/(1-\beta)>1$

3.Conclusion

Il est bien clair que l'utilisation du modèle de Markov pour prévoir la circulation future de livres particuliers n'est valable qu'en moyenne. Comme dans toutes les situations particulières, la pertinence varie considérablement d'un cas à l'autre ; certains livres dépassent la prévision, d'autres restent en deça. Cependant une politique basée sur cette évaluation est la meilleure qu'on puisse entreprendre, en moyenne, si l'on considère combien les présentes données ont un grand degré de variabilité.

Sauf si des informations supplémentaires relatives à l'usage futur d'un ouvrage particulier (son utilisation dans une classe, par exemple) s'avéraient utilisables, les prévisions tirées de l'utilisation du modèle constitue la meilleure base possible pour prendre une décision. Le modèle permet également d'évaluer quelles chances la circulation d'un livre a d'être supérieure ou inférieure aux prévisions, si cela est nécessaire à la prise de décisions.

On pourra trouver dans les écrits de Morse des calculs supplémentaires pour estimer l'intérêt d'acheter d'un exemplaire supplémentaire [MORS72, p.19] ou de mettre en réserve une partie du fonds [MORS68, Chap.8]

IV. Stationnarité

A. Les probabilités de transition

Si l'on désigne par X_t la variable qui représente le nombre de circulations pendant l'année t d'un groupe de livres donné, on sait, d'après l'hypothèse de Morse, que le nombre de circulation X_{t+1} pendant l'année t+1, sachant que $X_t = m$, suit une loi de Poisson de moyenne $\alpha+\beta m$. Et l'on notera

$$(X_{t+1} / X_t = m) \rightarrow \mathbf{P} (\alpha+\beta m)$$

La probabilité conditionnelle T_{mn}, définie dans la section III (équation 3.3), peut s'écrire à l'aide de ces notations:

T_{mn} = "*probabilité que* $X_{t+1}=n$ *sachant* $X_t=m$"

$$= P(X_{t+1}= n / X_t = m)$$

$$= \frac{(\alpha+\beta m)^n}{n!} e^{-(\alpha+\beta m)} \qquad (4.1)$$

et la moyenne conditionnelle N(m) définie en (3.6) s'écrira sous la forme

N(m)= "*Espérance de* X_{t+1} *sachant que* $X_t=m$"

$$= E(X_{t+1}/X_t = m)$$

$$= \sum_n n\, P(X_{t+1}= n / X_t = m)$$

$$= \alpha+\beta m \qquad (4.2)$$

A partir de l'équation (4.1), on peut prédire le comportement futur de la circulation. Par exemple, pour les livres qui ont eu m circulations pendant une année t donnée, la probabilité qu'ils aient n circulations pendant l'année t+1 suivante est T_{mn}, comme le donne l'équation (4.1); et la probabilité que l'un d'entre eux circule n fois durant l'année t+2 est

$$T^2_{mn} = T_{m0}T_{0n} + T_{m1}T_{1n} + T_{m2}T_{2n} + \ldots$$

$$= e^{-2\alpha}\left[\frac{\alpha^n}{n!} e^{-\beta m} + \frac{(\alpha+\beta m)(\alpha+\beta)^n}{1!\, n!} e^{-(m+1)\beta} + \frac{(\alpha+\beta m)^2(\alpha+\beta)^n}{2!\, n!} e^{-(m+1)\beta} \ldots \right]$$

Plus généralement, **la probabilité qu'un ouvrage circule n fois pendant l'année t_0+t sachant qu'il a circulé m fois pendant l'année t_0** est indépendante de t_0 et s'exprime à l'aide de la formule de Chapman-Kolmogorov de la façon suivante:

$$T^t_{mn} = P(X_t = n / X_{t0} = m)$$

$$T^t_{mn} = \sum_k T^{t-s}_{mk} T^s_{kn}$$
$$= T^{t-s}_{m0} T^s_{kn} + T^{t-s}_{mk} T^s_{kn} + T^{t-s}_{mk} T^s_{kn} + \ldots \qquad (4.3)$$

où s est un entier quelconque (0<s<t).La probabilité Tmn citée précédemment se notera T^1_{mn} dans la formule (4.3). Les éléments T^t_{mn} peuvent se ranger dans une matrice de transition notée $[T^t_{mn}]$.On trouvera dans l'appendice de *Library Effectiveness* [MORS68] les matrices $[T^t_{mn}]$ pour différentes valeurs de α, β et de t.

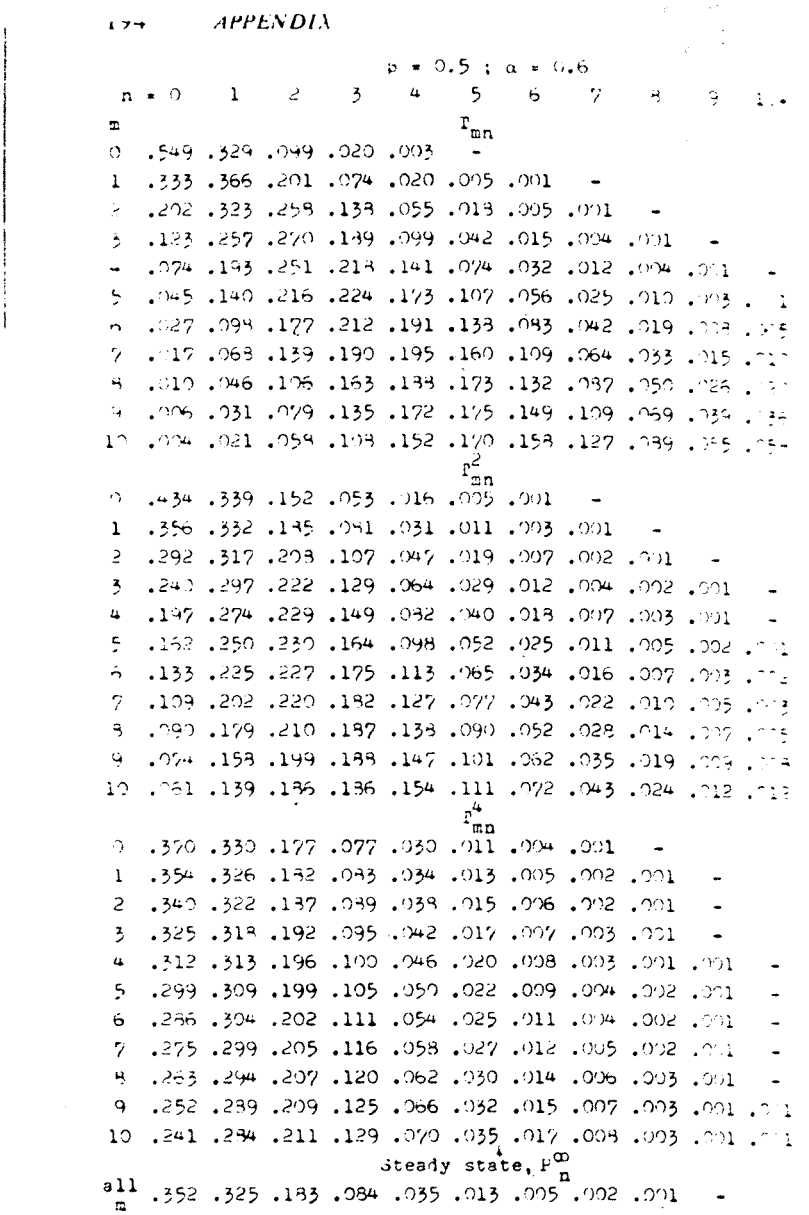
B.Stationnarité d'une collection

Une des propriétés inhérentes au processus de Markov est qu'à mesure que le temps s'écoule, une collection de livres tend à "oublier" quelle était sa circulation initiale.On peut constater (voir Tableau 4.1) que lorsque t augmente, les lignes de la matrice $[T^t_{mn}]$ sont de plus en plus semblables En d'autres termes la probabilité conditionnelle T^t_{mn} devient de moins en moins dépendante de la circulation initiale m lorsque t croît. Finalement, on a le résultat suivant:

$$\text{quand } t \to \infty, \qquad T^t_{mn} \to P_n^{\infty} \qquad (4.4)$$

ce qui signifie que la collection a atteint un état dit *stationnaire*. A ce moment, chaque ouvrage de la collection a la même *distribution de probabilité*, indépendamment de sa circulation initiale, ce qui se traduit par l'égalité des lignes

Library Effectiveness (cf Tableau 4.1). Elles représentent les distributions finales de la circulation des vieilles collections.



174 APPENDIX

p = 0.5 ; α = 0.6

m \ n	0	1	2	3	4	5	6	7	8	9	10
					r_{mn}						
0	.549	.329	.049	.020	.003	-					
1	.333	.366	.201	.074	.020	.005	.001	-			
2	.202	.323	.258	.138	.055	.018	.005	.001	-		
3	.123	.257	.270	.149	.099	.042	.015	.004	.001	-	
4	.074	.193	.251	.218	.141	.074	.032	.012	.004	.001	-
5	.045	.140	.216	.224	.173	.107	.056	.025	.010	.003	.1
6	.027	.098	.177	.212	.191	.133	.083	.042	.019	.008	.0
7	.017	.068	.139	.190	.195	.160	.109	.064	.033	.015	.0
8	.010	.046	.105	.163	.188	.173	.132	.087	.050	.026	.0
9	.006	.031	.079	.135	.172	.175	.149	.109	.069	.034	.0
10	.004	.021	.059	.108	.152	.170	.158	.127	.089	.055	.0
					r^2_{mn}						
0	.434	.339	.152	.053	.016	.005	.001	-			
1	.356	.332	.185	.081	.031	.011	.003	.001	-		
2	.292	.317	.208	.107	.047	.019	.007	.002	.001	-	
3	.240	.297	.222	.129	.064	.029	.012	.004	.002	.001	-
4	.197	.274	.229	.149	.082	.040	.018	.007	.003	.001	-
5	.162	.250	.230	.164	.098	.052	.025	.011	.005	.002	.0
6	.133	.225	.227	.175	.113	.065	.034	.016	.007	.003	.0
7	.109	.202	.220	.182	.127	.077	.043	.022	.010	.005	.0
8	.090	.179	.210	.187	.138	.090	.052	.028	.014	.007	.0
9	.074	.158	.199	.188	.147	.101	.062	.035	.019	.009	.0
10	.061	.139	.186	.186	.154	.111	.072	.043	.024	.012	.0
					r^4_{mn}						
0	.370	.330	.177	.077	.030	.011	.004	.001	-		
1	.354	.326	.182	.083	.034	.013	.005	.002	.001	-	
2	.340	.322	.187	.089	.038	.015	.006	.002	.001	-	
3	.325	.318	.192	.095	.042	.017	.007	.003	.001	-	
4	.312	.313	.196	.100	.046	.020	.008	.003	.001	.001	-
5	.299	.309	.199	.105	.050	.022	.009	.004	.002	.001	-
6	.286	.304	.202	.111	.054	.025	.011	.004	.002	.001	-
7	.275	.299	.205	.116	.058	.027	.012	.005	.002	.001	-
8	.263	.294	.207	.120	.062	.030	.014	.006	.003	.001	-
9	.252	.289	.209	.125	.066	.032	.015	.007	.003	.001	.0
10	.241	.284	.211	.129	.070	.035	.017	.008	.003	.001	.0
					steady state, p^{∞}_n						
all m	.352	.325	.183	.084	.035	.013	.005	.002	.001	-	

Tableau 4.1:Matrices [T^t_{mn}] , pour différentes valeurs de t, avec α et β fixés (extrait de l'appendice de *Library Effectiveness* [MORS68])

On se propose de démontrer cette propriété (4.4) de stationnarité en utilisant le résultat suivant [SAPO90, p.105] :

> On dit qu'un processus $\{X_t, t \geq 0\}$ est stationnaire *au sens strict* si sa loi de probabilité est invariante par translation sur t : ce qui signifie que $\{X_t\}$ et $\{X_{t+\tau}\}$, où $\tau \geq 0$ ont mêmes caractéristiques.
>
> La stationnarité au sens strict implique
>
> $$\mathbf{m_t = E(X_t) = m = constante} \qquad (1)$$
>
> et $$\mathbf{\sigma^2_t = Var(X_t) = \sigma^2 = constante} \qquad (2)$$
>
> ainsi que $$\mathbf{Cov(X_t, X_s) = \varphi(|t-s|)} \qquad (3)$$
>
> où φ est une fonction quelconque.
>
> Lorsque seules ces conditions sont remplies, on dit que le processus $\{X_t\}$ est stationnaire *au sens large.*

Si l'on veut appliquer ce résultat général au cas étudié, on doit montrer que lorsque $t \to \infty$ le processus $\{Xt\}$ du nombre de circulations est stationnaire au sens large. On va donc devoir démontrer que lorsque $t \to \infty$

a)	$n(t)=E(Xt) \to n(\infty)=$ constante
b)	$\sigma^2(t)=Var(Xt) \to \sigma^2_n=$ constante

Il n'est pas nécessaire de vérifier ici la condition (3) car elle est remplie par le critère d'homogénéité des processus markoviens[1].

On sera amenés à utiliser au cours de cette démonstration les deux théorèmes probabilistes généraux de l'espérance totale et de la variance totale:

Soit Y une variable aléatoire réelle, et X une autre variable aléatoire qui n'est pas nécessairement réelle mais peut être une variable qualitative. On peut alors définir, sous réserve de l'existence de ces expressions pour le cas dénombrable, l'espérance et la variance de Y à X fixé.

---

[1] Voir Annexe A "Généralités sur les processus markoviens"

⇨L'espérance conditionnelle

Définition

On appelle espérance de Y sachant que X=x et on note E(Y/X=x) la quantité définie par:

$$E(Y/X=x) = \sum_{y} yP(Y=y/X=x)$$

C'est donc l'espérance de Y prise par rapport à sa loi conditionnelle.On note que E(Y/X=x) est une fonction de x : $E(Y/X=x)=\varphi(x)$.

Définition

On appelle variable aléatoire "espérance conditionnelle de Y sachant X" et on note E(Y/X) la variable définie par :

$$E(Y/X) = \varphi(x)$$

Cette variable présente un certain nombre de propriétés remarquables (comme la linéarité), mais surtout on a, en prenant l'espérance de cette variable le

Théorème de l'espérance totale

$$E[\,E(Y/X)\,] = E(\cdot) \qquad (4.5)$$

⇨La variance conditionnelle

Définition

On appelle variance de Y sachant que X=x et on note Var(Y/X=x) la quantité:

$$Var(Y/X=x) = E[(Y-E(Y/X=x))^2 /X=x]$$

Comme pour l'espérance, et puisque $Var(Y/X=x) = \psi(x)$, on définit ensuite la variable aléatoire variance conditionnelle :

$$Var(Y/X) = \psi(x) = E[(Y - E(Y/X))^2 /X]$$

On a alors le résultat fondamental suivant

Théorème de la variance totale

$$Var(Y) = E[\ Var(Y/X)\ ] + Var[\ E(Y/X)\ ] \qquad (4.6)$$

■ Démontrons le résultat a) **$n(t)=E(Xt) \rightarrow n(\infty)=$ constante**

On suppose que le système démarre au temps t=0 dans l'état m ($X_0=m$). On va d'abord exprimer l'espérance conditionnelle suivante:

$$n_m(t) = E(X_t/\ X_0=m)$$

Sachant que $X_1/X_0=m \rightarrow P\ (\alpha+\beta m)$ on a donc

$$E(X_1/X_0=m) = \alpha+\beta m \qquad (a.1)$$

Comme $X_2/X_1=m \rightarrow P\ (\alpha+\beta m)$ on a

$$E(X_2\ /X_1=m) = \alpha+\beta m$$

d'où $E(X_2\ /X_1) = \alpha+\beta X_1$

et $E[\ X_2/\ (X_1/X_0=m)\ ] = \alpha+\beta(X_1/X_0=m) \qquad (a.2)$

Si l'on écrit la formule de l'espérance totale (4.5) avec $Y=X_t/X_0=m$ et $X=X_{t-1}/X_0=m$ (ce sont ici des variables conditionnées), cela donne

$$E(X_t/\ X_0=m)=E\{\ E[X_t/\ (X_{t-1}/X_0=m)]\ \} \qquad \forall t>1$$

En écrivant cette formule pour t=2, on obtient

$$E(X_2/ X_0=m) = E\{ E[X_2 / (X_1/X_0=m)] \}$$

$$= E[\alpha+\beta(X_1/X_0=m)] \qquad \text{(d'après a.2)}$$

$$= \alpha+\beta\ E(X_1/ X_0=m)$$

$$= \alpha+\beta\ (\alpha+\beta m) \qquad \text{(d'après a.1)}$$

$$= \alpha(1+\beta) + \beta^2\ m$$

A l'aide d'un raisonnement similaire, on obtiendra

$$E(X_3 /X_0=m) = \alpha(1+\beta+ \beta^2) + \beta^3\ m$$

Par itérations successives on obtient la formule générale

$$E(X_t /X_0=m) = \alpha(1+\beta+ \beta^2+ \ldots +\beta^{t-1}\ ) + \beta^t\ m$$

soit

$$n_m(t)= \alpha\ \frac{(1-\beta^t)}{(1-\beta)} + \beta^t m \qquad (4.7)$$

On constate que lorsque $t\to\infty$, la variance conditionnelle $n_m(t)$ tend vers $\alpha/(1-\beta)$. A l'aide de (4.7) on peut maintenant exprimer l'espérance totale

$$n(t) = E(X_t)$$

A l'aide de la formule de l'espérance totale (4.5) on peut écrire la relation

$$E(X_t) = E[\ E(X_t/X_0)\ ]$$

$$= E\left[\ \alpha\ \frac{(1-\beta^t)}{(1-\beta)} + X_0\ \beta^t\ \right] \qquad \text{(d'après 4.7)}$$

$$= \alpha/(1-\beta)\ (1-\beta^t) + E(X_0)\beta^t$$

En posant $M= E(X_0)$ et $n(\infty)=\alpha/(1-\beta)$, on obtient

$$n(t) = n(\infty)(1-\beta^t) + M\beta^t$$

$$\boxed{n(t) = n(\infty) + [M-n(\infty)]\beta^t \qquad (4.8)}$$

De (4.8) on déduit que si $\beta<1$ alors $n(t)\rightarrow n(\infty)$ lorsque $t\rightarrow\infty$, ce qui est bien ce que nous voulions démontrer.

■Démontrons maintenant le résultat b) $\sigma^2(t)=\mathbf{Var(X_t)}\rightarrow\sigma^2(\infty)$ **=constante** :

Comme $X_1/X_0=m \rightarrow P(\alpha+\beta m)$ et que la variance d'une loi de Poisson est égale à sa moyenne, on peut écrire

$$Var(X_1/X_0=m) = E(X_1/X_0=m) = \alpha+\beta m$$

On écrit ensuite la formule de la variance totale (4.6) avec $Y=X_t/X_0=m$ et $X=X_{t-1}/X_0=m$; cela donne :

$$Var(X_t/X_0=m) = Var\{E[X_t/(X_{t-1}/X_0=m)]\} + E\{Var[X_t/(X_{t-1}/X_0=m)]\} \qquad (4.9)$$

Pour t=2, on obtient

$$Var(X_2/X_0=m) = Var\{E[X_2/(X_1/X_0=m)]\} + E\{Var[X_2/(X_1/X_0=m)]\}$$

$$= Var\{\ \alpha+\beta(X_1/X_0=m)\ \} + E\{\ \alpha+\beta(X_1/X_0=m)\ \}$$

$$= \beta^2\ Var(X_1/X_0=m) + \alpha + \beta\ E(X_1/X_0=m)$$

$$= \beta^2\ (\alpha+\beta m) + \alpha + \beta(\alpha+\beta m)$$

$$= \alpha(1+\beta+\beta^2) + m\ \beta^2\ (1+\beta)$$

Pour t=3, la formule de la variance totale s'écrit

$$\mathrm{Var}(X_3/X_0=m) = \mathrm{Var}\{E[X_3/(X_2/X_0=m)]\} + E\{\mathrm{Var}[X_2/(X_2/X_0=m)]\}$$

$$= \mathrm{Var}\{\ \alpha+\beta(X_2/X_0=m)\ \} + E\{\ \alpha+\beta(X_2/X_0=m)\ \}$$

$$= \beta^2\ \mathrm{Var}(X_2/X_0=m) + \alpha + \beta\ E(X_2/X_0=m)$$

$$= \beta^2\ [\alpha(1+\beta+\beta^2) + m\,\beta^2\,(1+\beta)] + \alpha + \beta[\alpha(1+\beta) + \beta^2\,m]$$

$$= \alpha\beta^2\,(1+\beta+\beta^2) + \alpha\,(1+\beta+\beta^2) + m\,(\beta^3+\beta^4+\beta^5)$$

$$= \alpha\,(1+\beta+\beta^2)\,(1+\beta^2) + m\beta^3\,(1+\beta+\beta^2)$$

$$= \alpha\ \frac{(1-\beta^3)}{(1-\beta)}\ \frac{(1-\beta^4)}{(1-\beta^2)} + m\beta^3\ \frac{(1-\beta^3)}{(1-\beta)}$$

Démontrons par récurrence que l'on a l'expression suivante

$$\mathrm{Var}(X_t/X_0=m) = \alpha\ \frac{(1-\beta^t)}{(1-\beta)}\ \frac{(1-\beta^{t+1})}{(1-\beta^2)} + m\beta^t\ \frac{(1-\beta^t)}{(1-\beta)} \qquad \forall t\geq 3 \qquad (4.10)$$

Cette expression est vraie aux rangs t = 3 et t = 4.

On fait l'hypothèse que la formule (4.10) est vraie au rang t .Démontrons qu'elle est vraie au rang t+1. La variance de X_{t+1} sachant que $X_0=m$ s'écrit sous l'hypothèse de la relation (4.10)

$$\mathrm{Var}(X_{t+1}/X_0=m)$$

$$= \mathrm{Var}\{E[X_{t+1}/(X_t/X_0=m)]\ \} + E\{\mathrm{Var}[X_{t+1}/(X_t/X_0=m)]\} \quad \text{(d'après 4.9)}$$

$$= \beta^2\ \mathrm{Var}(X_t/X_0=m) + \alpha + \beta\ E(X_t/X_0=m)$$

$$= \beta^2\ \mathrm{Var}(X_t/X_0=m) + \alpha + \beta\ \left[\alpha\ \frac{(1-\beta^t)}{(1-\beta)}\right] + \beta^t\,m \quad \text{(d'après 4.7)}$$

$$= \beta^2\ \left[\alpha\ \frac{(1-\beta^t)}{(1-\beta)}\ \frac{(1-\beta^{t+1})}{(1-\beta^2)} + m\beta^t\ \frac{(1-\beta^t)}{(1-\beta)}\right] + \alpha + \beta\alpha\ \frac{(1-\beta^t)}{(1-\beta)} + \beta^{t+1}m$$

d'après l'hypothèse (4.10). On obtient finalement

$$\mathrm{Var}(X_{t+1} / X_0=m)$$

$$= \alpha \frac{(\beta^2 - \beta^{t+2})(1 - \beta^{t+1})}{(1-\beta)(1-\beta^2)} + \alpha \left[1 + \frac{\beta(1-\beta^t)}{(1-\beta)} \right] + m\beta^{t+1} \left[1 + \frac{\beta(1-\beta^t)}{(1-\beta)} \right]$$

$$= \alpha \frac{(\beta^2 - \beta^{t+2})(1 - \beta^{t+1})}{(1-\beta)(1-\beta^2)} + \alpha \frac{(1-\beta^{t+1})}{(1-\beta)} + m\beta^{t+1} \frac{(1-\beta^{t+1})}{(1-\beta)}$$

$$= \alpha \frac{(1-\beta^{t+1})}{(1-\beta)} \left[1 + \frac{(\beta^2-\beta^{t+2})}{(1-\beta^2)} \right] + m\beta^{t+1} \frac{(1-\beta^{t+1})}{(1-\beta)}$$

$$= \alpha \frac{(1-\beta^{t+1})}{(1-\beta)} \frac{(1-\beta^{t+2})}{(1-\beta^2)} + m\beta^{t+1} \frac{(1-\beta^{t+1})}{(1-\beta)}$$

On vient donc de démontrer que si l'expression 4.10 est vraie au rang t alors elle est vraie au rang t+1 . Par conséquent, pour tout entier $t > 2$ on a

$$\sigma^2_m(t) = \mathrm{Var}(X_t / X_0=m) = \alpha \frac{(1-\beta^t)}{(1-\beta)} \frac{(1-\beta^{t+1})}{(1-\beta^2)} + m\beta^t \frac{(1-\beta^t)}{(1-\beta)}$$

On peut maintenant, à l'aide de la formule (4.10) exprimer la variance totale

$$\sigma^2(t) = \mathrm{Var}(X_t)$$

Avec la formule de la variance totale (4.6), on obtient

$$\mathrm{Var}(X_t) = \mathrm{Var}[\, E(X_t/X_0) \,] + E[\, \mathrm{Var}(X_t/X_0) \,]$$

$$= \mathrm{Var}\left[\alpha \frac{(1-\beta^t)}{(1-\beta)} + X_0\beta^t \right] + E\left[\alpha \frac{(1-\beta^t)}{(1-\beta)} \frac{(1-\beta^{t+1})}{(1-\beta^2)} + m\beta^t \frac{(1-\beta^t)}{(1-\beta)} \right]$$

Finalement

$$\sigma^2(t) = \beta^{2t}\,\mathrm{Var}(X_0) + \alpha\,\frac{(1-\beta^t)(1-\beta^{t+1})}{(1-\beta)(1-\beta^2)} + \beta^t\,\frac{(1-\beta^t)}{(1-\beta)}\,E(X_0)$$

Donc si $\beta < 1$ on a $\sigma^2(t) \to \alpha/(1-\beta)(1-\beta^2) = \sigma^2(\infty)$

V. Etude des paramètres α et β

A. *Méthodes de calcul et propriétés*

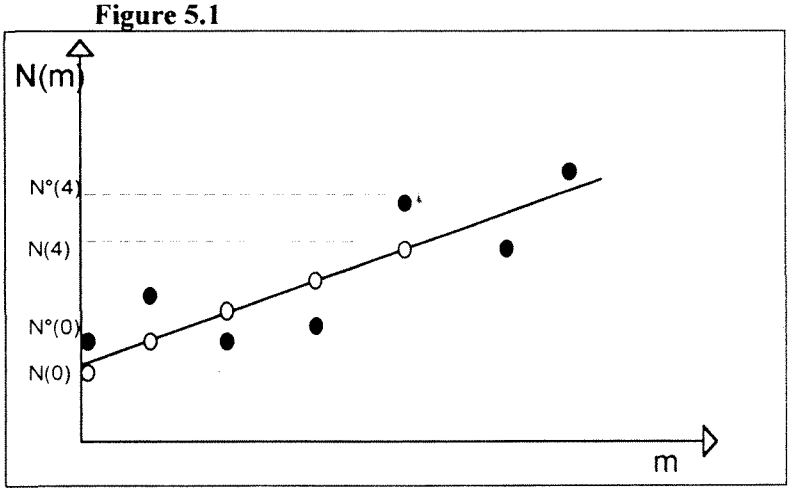
1. Méthodes de calcul

Nous avons vu dans la section III que si l'on trace sur un graphe les valeurs expérimentales N°(m) (nombre moyen de circulations d'un ouvrage durant l'année t+1 sachant qu'il a circulé m fois durant l'année t) en fonction de m, on constate que les points (m, N°(m)) sont pratiquement alignés suivant une droite d'équation.

$$\mathbf{N(m)=\alpha+\beta m} \qquad (5.1)$$

N(m) étant la valeur théorique de N°(m), obtenue avec l'hypothèse de cette relation linéaire (5.1).
Il reste à estimer les coefficients αet β afin que la droite d'équation (5.1) ajuste au mieux le nuage de points (Fig 5.1).

Figure 5.1



⇨ **Méthode empirique de Morse:**

Pour calculer α et β pour une classe de livres étudiée, on prend la somme des circulations observées dans l'échantillon pour deux années successives et on la divise par la circulation de la première année. On calcule ainsi la circulation moyenne durant l'année écoulée de tous les livres ayant circulé m fois l'année précédente, $R_m(t)$ (ce qui correspond à N°(m)).

Par exemple, si 20 livres de l'échantillon ont circulé exactement 2 fois durant l'année t-1 et si la circulation totale de ces 20 livres durant l'année écoulée t a été 32, $R_2(t)$ pour cet échantillon sera 32/20 = 1,6;et si les 80 livres restant, qui n'ont pas circulé durant l'année t-1(les livres inactifs de cette année-là) ont eu une circulation de 32 durant l'année t, $R_0(t)$ sera 32/80 = 0,4. On note qu'un livre inactif une année peut être actif l'année suivante, ce qui est souvent le cas dans la réalité, ce qui est pris en compte dans la formule de Markov.

On calcule $R_0(t)$, $R_1(t)$, $R_2(t)$, $R_3(t)$ et $R_4(t)$ à partir des données de l'échantillon. Et pour calculer α et β on utilisera les équations suivantes

$$\alpha = R_0(t)$$
$$\beta = 1/10\ [R_1(t)+R_2(t)+R_3(t)+R_4(t)-4R_0(t)] \qquad (5.2)$$

Dans l'équation $N(m)=\alpha+\beta m$, α représente l'ordonnée à l'origine. Remarquons que Morse a choisi de prendre $\alpha=R_0(t)=N°(0)$ alors que la meilleure droite de régression n'est pas forcément celle qui passe par le point (0,N°(0)). Mais ce choix semble justifié par le fait que les ouvrages qui n'ont pas circulé durant la première année t-1 ont un poids très important.

Cette méthode peut cependant présenter des inconvénients: il peut être en effet difficile de calculer $R_0(t)$, $R_1(t)$, $R_2(t)$, $R_3(t)$ et $R_4(t)$ pour les catégories de livres qui ont un fort taux de rotation, faute d'avoir trouvé un nombre suffisant de volumes ayant été empruntés 0, 1, 2, 3 et 4 fois l'avant-dernière année.C'est ce que fait remarquer Cane [CANE87]:

> *"Nous avons rencontré ce problème en essayant de faire l'étude des bandes dessinées; sans doute nos données portaient-elles sur un nombre très réduit de livres, mais même si nous avions pu effectuer la totalité du travail prévu sur cette catégorie, nous aurions dû calculer α et β à partir de R_0, R_1, R_2, R_3 et R_4 issus de données portant sur moins d'une demi-douzaine d'ouvrages, ce qui ne paraît pas une méthode fiable.*

Deux solutions paraissent possibles : soit augmenter la taille de l'échantillon dans les catégories où le taux de rotation est particulièrement fort, ce qui reste simple, soit à déterminer les valeurs de α et β à partir de la méthode des moindres carrés."

⇨ **Méthode des moindres carrés**

On cherche à ajuster au nuage de points (m,N°(m)) une droite d'équation $N(m)=\alpha+\beta m$ de sorte que

$$\sum_{m=0}^{M}(N°(m) - N(m))^2$$

soit minimale. On cherche donc à minimiser la fonction

$$F(\alpha,\beta) = \sum_{m\geq 0}(N°(m)-\alpha-\beta m)^2$$

Ce minimum est atteint pour $\partial F/\partial\alpha = \partial F/\partial\beta = 0$, ce qui donne deux équations:

$$\sum_{m=0}^{M}(N°(m)-\alpha-\beta m) = 0$$

et

$$\sum_{m=0}^{M} m(N°(m)-\alpha-\beta m) = 0$$

dont les solutions sont

$$\beta = 1/\sum_{m\geq 0}(m-m_{moy})\left(\sum(m-m_{moy})\, N°(m)\right)$$

et (5.2)

$$\alpha = N°_{moy} - \beta\, m_{moy}$$

avec $m_{moy} = (1/M) \sum_{m \geq 0} m$

et $N°_{moy} = (1/M) \sum_{m \geq 0} N°(m)$

⇨ Méthode des moindres carrés pondérés

Si X_m est le nombre de circulations pour une année donnée, des ouvrages ayant circulé m fois l'année précédente, $P_m = X_m / \sum X_m$ représente la fréquence de circulation des ouvrages qui sont sortis m ($m \geq 0$) fois l'année précédente. Il semble naturel, pour calculer α et β de pondérer les couples $(m, N°(m))$ par P_m; en effet, la contribution de $N°(m)$ est plus significative d'un point de vue statistique pour les petites valeurs de m.

Dans le cas d'une régression linéaire pondérée, on cherche à minimiser l'écart

$$G(\alpha,\beta) = \sum_{m=0}^{M} P_m \left(N°(m) - \alpha - \beta m\right)^2$$

où le triplet $(m, N°(m), P_m)$, avec $0 \leq m \leq M$, désigne respectivement l'abscisse, l'ordonnée et le poids. On a $\sum P_m = 1$. On obtient comme valeurs de α et β

$$\beta = \frac{S_1 - M' N'_{moy}}{S_2 - (N'_{moy})^2} \qquad (5.3)$$

$$\alpha = N'_{moy} - \beta M'$$

avec

$$M' = \sum_{m=0}^{M} m P_m, \qquad S_1 = \sum_{m=0}^{M} m\, N°(m) P_m$$

$$N'_{moy} = \sum_{m=0}^{M} N°(m) P_m, \qquad S_2 = \sum_{m=0}^{M} m^2 P_m$$

Si Pm=1/M, on retrouve les valeurs des coefficients de régression linéaire non pondérée.

2.Propriétés barycentriques des coefficients α et β

Dans le cas où plusieurs sous-classes d'une même classe de livres ont été observées durant des périodes différentes, les valeurs α et β de la classe totale peuvent être aisément calculées en utilisant les équations suivantes

$$\alpha = \frac{M1\alpha1+M2\alpha2+\ldots}{M1+M2+\ldots}$$

$$\beta = \frac{M1\beta1+M2\beta2+\ldots}{M1+M2+\ldots} \qquad (5.4)$$

où M1 livres d'une sous-classe ont des paramètres $\alpha1$ et $\beta1$ et M2 livres d'une deuxième sous-classe ont des paramètres de valeurs $\alpha2$ et $\beta2$.

Cette propriété est due au caractère linéaire du modèle de l'équation (5.1), $N(m)=\alpha+\beta m$: puisque l'équation (5.1) est linéaire en α, β et m, les valeurs de α et β peuvent être obtenues à l'aide d'équations de même forme.

■ On peut faire la démonstration de ce résultat dans le cas le plus simple où une catégorie A de livres, de paramètres α et β, est scindée en deux sous-catégories A1 et A2 de paramètres respectifs ($\alpha1$, $\beta1$) et ($\alpha2$, $\beta2$).

On définit les quantités suivantes:

- $M_1=\sum M_1(m)$ est le nombre d'ouvrages de A1, où $M_1(m)$ est le nombre d'ouvrages ayant circulé m fois durant une année t donnée.

- $M_2=\sum M_2(m)$ est le nombre d'ouvrages de A2 qui ont circulé m fois pendant une année t.

Notons également que $M(m)=M_1(m)+M_2(m)$ est le nombre d'ouvrages de la catégorie A qui ont circulé m fois durant l'année t. Et $M=M_1+M_2$ est le nombre total d'ouvrages de la catégorie A.

On peut écrire la relation (5.1) qui nous donne la circulation moyenne N(m) durant l'année t+1, pour la catégorie A et les deux sous-catégories A1 et A2:

A: $N(m)=\alpha+\beta m$ (1)

A2: $N_1(m)=\alpha 1+\beta 1 m$ (2)

A3: $M_2(m)=\alpha 2+\beta 2 m$ (3)

La circulation moyenne conditionnelle N(m) de la catégorie A peut s'exprimer en fonction des paramètres de A1 et de A2 de la façon suivante

$$N(m) = \frac{M1(m)N1(m)+M2(m)N2(m)}{M1(m)+M2(m)}$$

ce qui s'écrit encore en remplaçant N(m), N1(m) et N2(m) par leurs expressions (1), (2) et (3)

$$\alpha+\beta m = \frac{M1(m)\ (\alpha 1+\beta 1 m)+M2(m)\ (\alpha 2+\beta 2 m)}{M1(m)+M2(m)}$$

$$\alpha+\beta m = \frac{\alpha 1 M1(m)+\alpha 2 M2(m)}{M1(m)+M2(m)} + m\,\frac{\beta 1 M1(m)+\beta 2 M2(m)}{M1(m)+M2(m)}$$

On en déduit par identification

$$\alpha = \frac{\alpha 1 M1(m)+\alpha 2 M2(m)}{M1(m)+M2(m)} \qquad (4)$$

$$\beta = \frac{\beta 1 M1(m)+\beta 2 M2(m)}{M1(m)+M2(m)} \qquad (5)$$

De la relation (4), on tire

$$\alpha\,(M1(m)+M2(m)) = \alpha 1 M1(m)+M2(m) \qquad \forall m$$

En effectuant une sommation sur les m, on obtient

$$\alpha \sum_{m\geq 0}(M1(m)+M2(m)) = \alpha 1 \sum_{m\geq 0} M1(m)+\alpha 2 \sum_{m\geq 0} M2(m)$$

$$\alpha(M1+M2) = \alpha 1 M1+\alpha 2 M2$$

On tire de cette dernière relation

$$\alpha = \frac{\alpha 1 M1 + \alpha 2 M2}{(M1+M2)}$$

Par le même raisonnement on obtient

$$\beta = \frac{\beta 1 M1 + \beta 2 M2}{M1+M2}$$

■

B. Dépendance temporelle des coefficients α et β

Rappelons qu'en ce qui concerne l'évolution dans le temps des paramètres α et β, Morse faisait les hypothèses suivantes:

1. Le paramètre β qui mesure la diminution de "popularité" d'un livre d'une année sur l'autre reste en gros constant pendant 10 à 12 ans de vie du livre à la bibliothèque

2. Le paramètre α, qui mesure la circulation moyenne atteinte par les ouvrages les plus anciens, est indépendant du temps.

Morse suggère que α tombe aux 2/3 de sa valeur après une période de 10 à 12 ans, tandis que Chen note que "bien que α est plus ou moins indépendant du temps, il se change généralement au bout de T années en α', où α' est généralement inférieur à α ".

Des chercheurs comme Kraft [KRAF70] ou Beshesti et Tague [BESH84] se sont intéressés à travers leurs études portant sur le modèle de Morse aux fluctuations de ces deux paramètres.

1. Le modèle de Morse revu par Beshesti et Tague

Beshesti et Tague ont testé le modèle de Morse à la bibliothèque "Library of Congress" de l'Université du Saskatchewan (Canada). Ils ont observé la circulation relative à 56 040 monographies de l'année 1967-1968 à 1977-1978. Ils ont enregistré pendant ces 11 années 99 430 transactions, ce qui représente un ensemble de données très large (Chen, dans sa thèse, utilisait un échantillon de 12 000 transactions).

Puisque Morse suggérait que son modèle était approprié pour prédire la circulation moyenne d'une classe de livres, Beshesti et Tague ont défini trois classes "Q", "D" et "N", qui représentent respectivement les sciences, l'histoire (sciences sociales) et les arts.

Pour tester la validité du modèle , 40 tables ont été construites sur les 11 années d'étude, 10 pour la collection considérée dans son intégralité, et 10 pour chacune des trois classes.Chaque table contient les données suivantes:

M(m): nombre de monographies ayant circulé m fois pendant l'année t

N(m,n) : nombre de monographies ayant circulé m fois pendant l'année t et n fois pendant l'année t+1.

N(m): nombre moyen de circulations pendant l'année t+1 des monographies qui ont circulé m fois pendant l'année t.

2.L'approximation linéaire pour 99% des données

La figure 5.2 montre qu'il existe une dépendance linéaire entre N(m) et m pour à peu près la moitié des observations. L'approximation linéaire est satisfaisante uniquement pour les livres qui n'ont pas circulé fréquemment (m≤8) l'année précédente

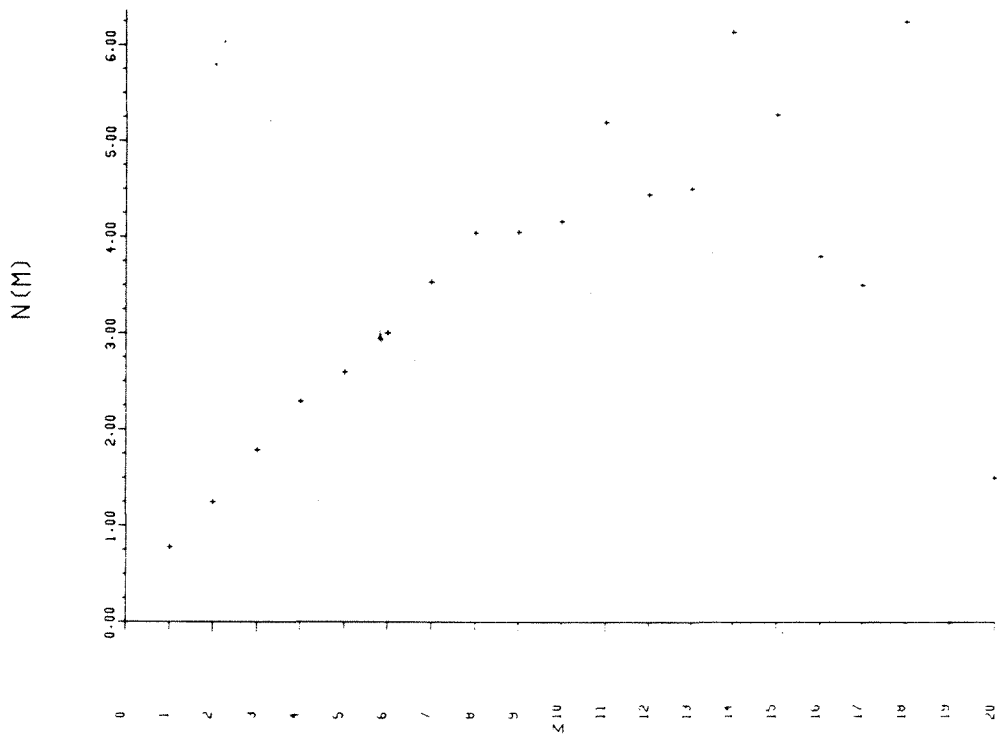


Figure 5.2 Circulation moyenne N(m) pour l'année 1968-1969 en fonction de la circulation m de l'année 1967-1968 [BESH 84]

En fait, le coefficient de détermination de la régression,R^2, entre les deux variables N(m) et m pour le premier couple d'années est inférieur à 0,35, ce qui indique une corrélation d'approximativement R=0,6 entre le nombre de circulations des monographies en 1967-1968 et leur nombre moyen de circulation pendant l'année suivante.

Le tableau (5.1) donne les valeurs du R^2 pour les 10 couples d'années, et cela pour la collection totale et pour les trois classes. Pour chaque classe, la colonne de droite, sous le titre 99, donne les valeurs du R^2 lorsque seulement 99,5% des données ont été utilisées pour construire les tables, c'est-à-dire quand les ouvrages qui circulaient les plus fréquemment -ce qui contribuait à moins de 0,5% du nombre total de transactions- ont été éliminés des calculs. La plupart des valeurs du R^2 sont considérablement plus élevées quand seulement 99,5% des données sont prises en compte. Beshesti et Tague en concluent que les ouvrages qui ont circulé plus de 8 fois pendant l'année 67-68 forment une "longue queue" dans la distribution de la circulation des documents, et par conséquent réduisent son adéquation au modèle linéaire.

Il est possible que Morse et Chen, qui ont travaillé sur des ensembles de données plus restreints n'aient point remarqué cette caractéristique de la distribution, puisque qu'environ 99% des données satisfont au modèle linéaire.

Tableau 5.1 : Coefficient de corrélation R^2 entre le nombre de transactions de l'année indiquée et celui de l'année précédente [BESH84]

	All		Q		D		N	
Year	100	99	100	99	100	99	100	99
1	0.309	0.987	0.025	0.988	−0.037	0.969	0.696	0.845
2	0.784	0.978	0.603	0.950	0.604	0.972	0.091	0.878
3	0.612	0.903	0.327	0.773	0.497	0.826	0.041	0.967
4	0.465	0.963	0.030	0.797	0.757	0.919	0.568	0.498
5	0.247	0.950	0.063	0.882	0.583	0.946	0.806	0.955
6	0.469	0.979	0.552	0.799	0.425	0.897	0.108	0.730
7	0.638	0.979	0.628	0.924	0.367	0.791	0.548	0.612
8	0.191	0.975	0.809	0.642	0.781	0.971	0.264	0.587
9	0.135	0.991	0.959	0.902	0.643	0.982	0.881	0.883
10	0.783	0.983	0.803	0.916	0.967	0.926	0.115	0.945

Pour tester une éventuelle dépendance temporelle du paramètre α, on a relevé les valeurs de α pour chaque couple d'années (Figures 5.3-5.6). Comme le tableau 5.2 l'indique, les coefficients de corrélation entre la variable α et la variable temps t, se situent entre -0,778 et -0,939. En fait α **décroît linéairement avec le**

temps, comme on peut le constater sur les figures 5.3-5.6, et par conséquent α est dépendant du temps.

Il est intéressant de remarquer que pour la collection intégrale et les classes D et Q, α croît entre la première et la seconde année pour diminuer ensuite. Peut-être que le déclin ne commence-t'il vraiment qu'à partir de la troisième année. Mais quelque soit la forme de ce déclin les courbes de α en fonction du temps ont une allure décroissante, et ceci queqlue soit la classe considérée.

D'un autre côté, β semble montrer quelque dépendance temporelle pour la classe Q, et dans une moindre mesure pour la classe N (cf. tableau 5.2), ce qui indique que le taux avec lequel ces deux classes perdent leur popularité est dépendant du temps, tandis que la collection totale et la classe D deviennent impopulaires avec un taux plus ou moins aléatoire. Il apparaît cependant qu'en général les fluctuations du paramètre β avec le temps sont plus ou moins aléatoires et ne permettent pas de conclure à une quelconque dépendance temporelle de ce coefficient.

Be.shestî et Tague soulignent qu'en tous les cas **ni α ni β ne peuvent être utilisés pour calculer les valeurs asymptotiques de la circulation** comme le suggèrait Morse (cf Chap.III).

Tableau 5.2 : Coefficients de corrélation entre les valeurs de α et β et le temps, pour la collection totale et les trois classes. [BESH 84]

Collection	*a* *r*	*b* *r*
All	−0.917	−0.037
Q	−0.778	−0.614
D	−0.939	−0.034
N	−0.836	−0.409

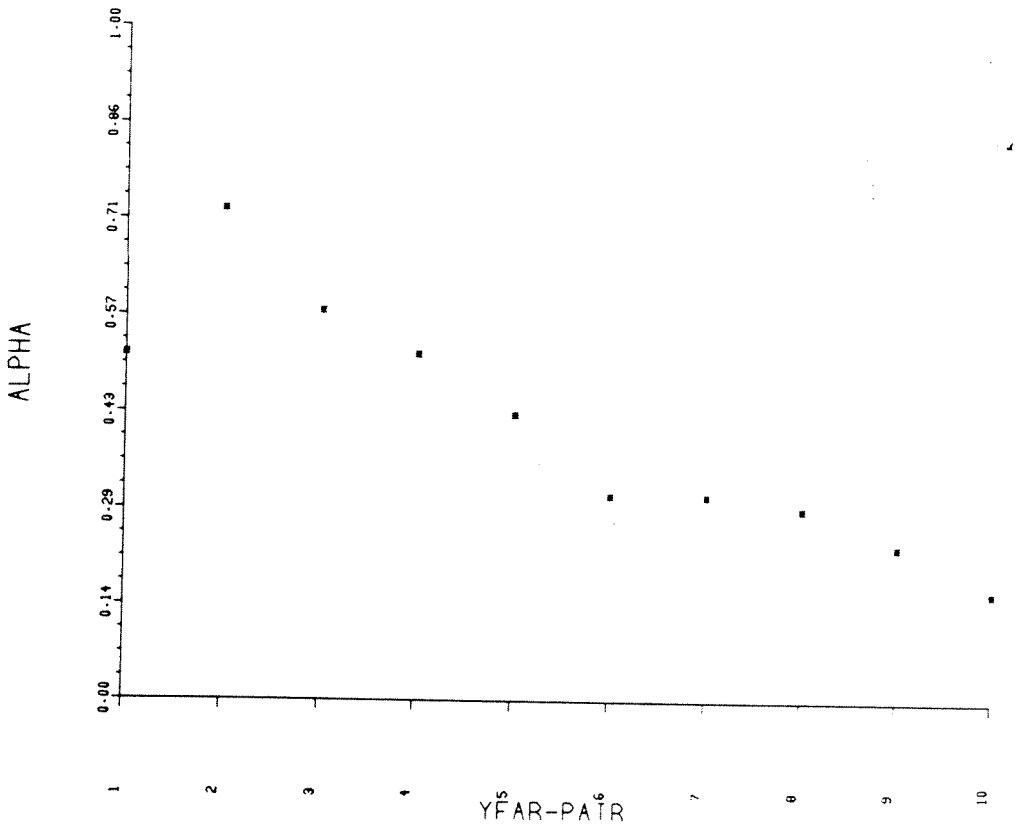


Figure 5.3 : α en fonction du temps pour la collection globale [BESH84]

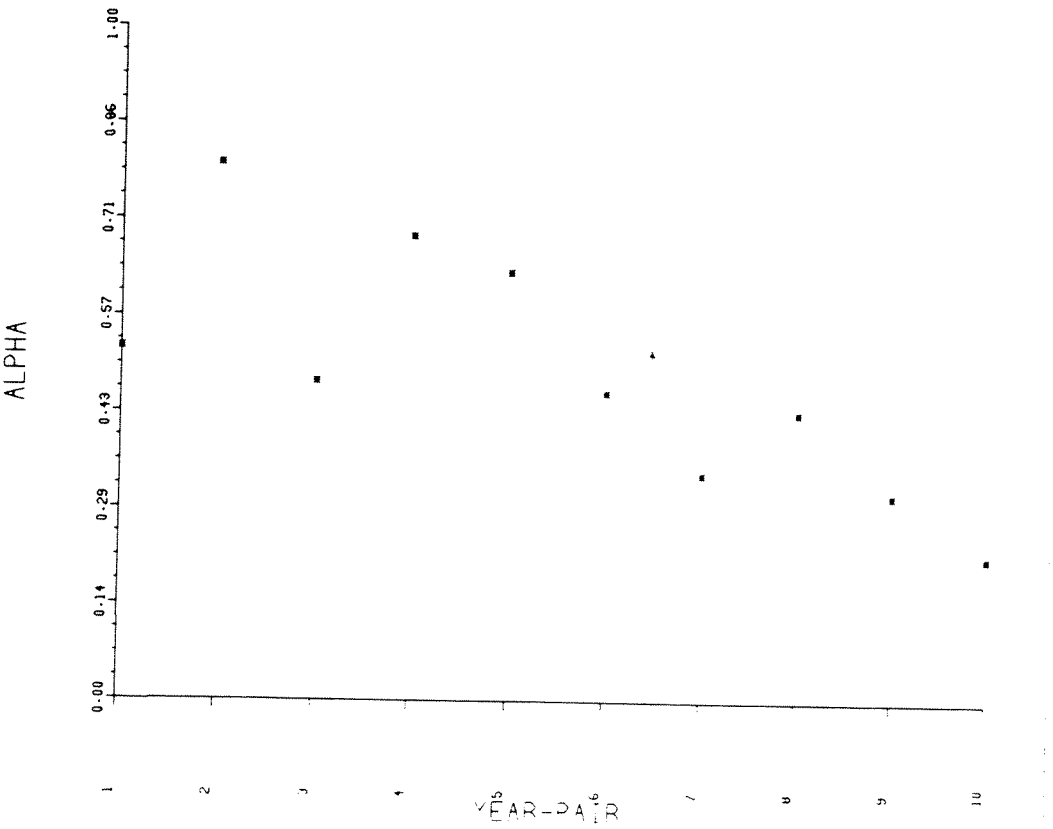


Figure 5.4 : α en fonction du temps pour la classe "Q" [BESH84]

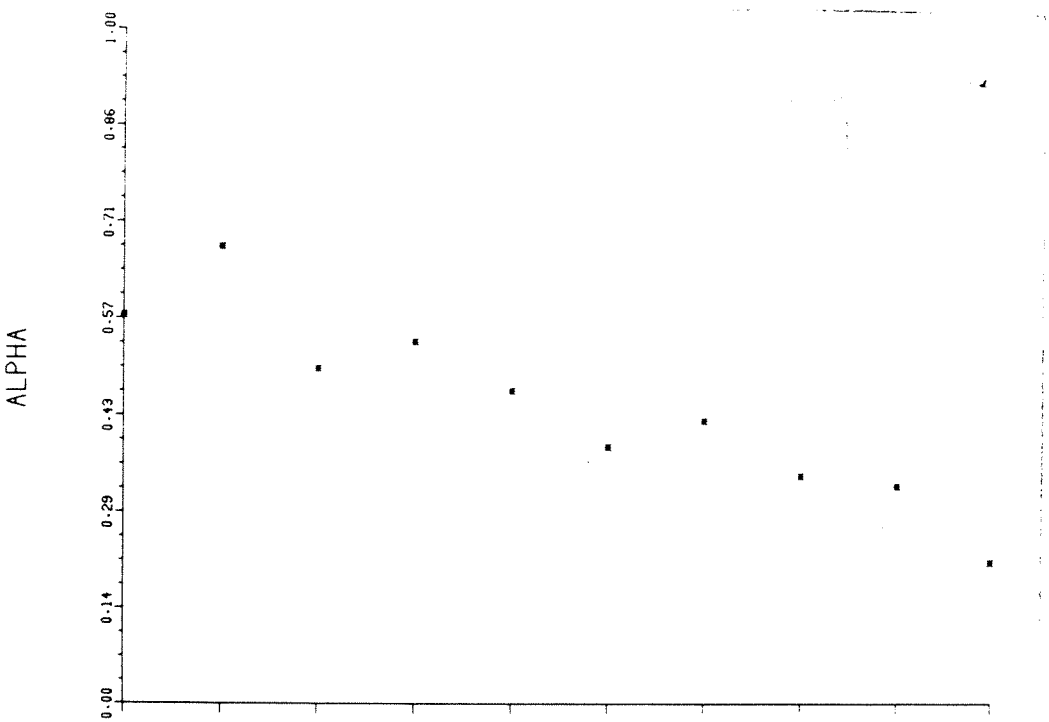


Figure 5.5 : α en fonction du temps pour la classe "D" [BESH84]

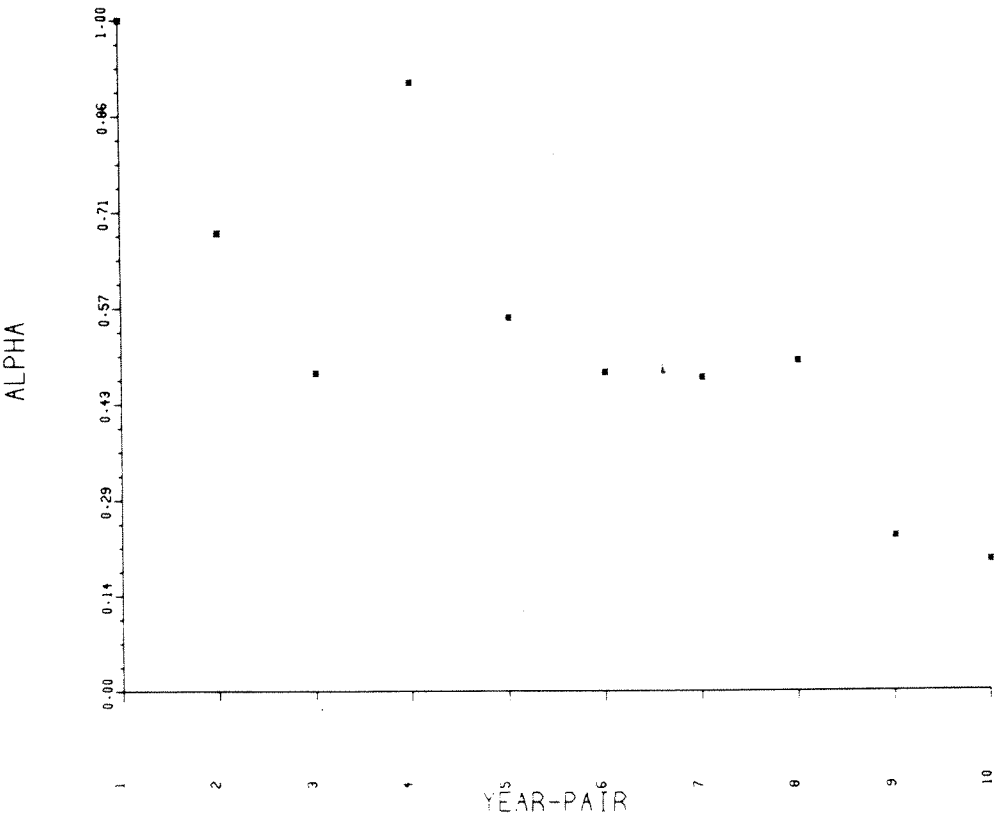


Figure 5.6 : α en fonction du temps pour la classe "N" [BESH84]

3 Le nombre de transactions par an

Une autre variable qui n'a pas été prise en compte dans le modèle de Morse original est le nombre de transactions total par an. Si la population d'une université reste relativement stable, cette variable peut refléter les modifications du comportement des usagers (en fait la meilleure façon d'évaluer le comportement des utilisateurs serait de connaître nombre moyen annuel de transactions par document, mais ces données n'étaient pas disponibles dans le cadre de cette étude) .

Pour les classes Q et N, les coefficients de corrélation entre le nombre total de transactions par an et les valeurs de β sont plus élevés qu'avec les valeurs de α (cf. tableau 5.3), ce qui indique que le taux avec lequel ces classes perdent leur popularité est plus dépendant des habitudes des usagers qu'il ne le serait pour la collection totale ou la classe D. En général, les coefficients de corrélation sont suffisamment élevés pour justifier l'introduction de la variable s, "nombre de transactions par an", dans le modèle.

Tableau 5.3 : Coefficients de corrélation entre α et et β et le nombre total de transactions par an, pour la collection globale et les trois classes. [BESH84]

Collection	α r	β r
All	0.917	−0.082
Q	0.677	0.688
D	0.813	0.066
N	0.465	0.543

Finalement, afin de tester formellement l'hypothèse d'une dépendance temporelle du paramètre α, Beshesti et Tague proposent une version modifiée du modèle de Morse, qui inclut explicitement le temps. Si α est dépendant linéairement du temps alors la moyenne conditionnelle du nombre de circulations peut être approchée par la nouvelle relation linéaire

$$N(m,t)=\alpha_t+\beta m$$

où α_t est un fonction linéaire décroissante du temps, et peut donc s'écrire

$\alpha_t=\alpha+ct \qquad (c<0)$

Le nouveau modèle linéaire sera donc

(2) $N(m,t)=\alpha+\beta m+ct$

où α et β et c sont des constantes, et t est la variable qui représente le temps.Il s'agit d'une régression linéaire multiple à 2 variables explicatives m et t.

Pour tester l'influence sur le modèle du comportement des usagers dans leurs activités d'emprunts en termes de nombre de transactions par an, le modèle de Morse peut être modifié en incluant explicitement ce volume de transactions

(3) $N(m,s)=\alpha+\beta m+ds$

où α et β et d sont des constantes et s est le nombre total de transactions.

Le modèle peut aussi s'écrire sous la forme d'une régression linéaire multiple qui prend en compte les trois variables explicatives m, t, s:

(4) $N(m,t,s)=\alpha+\beta m+ct+ds$

Les coefficients de ces différents modèles ont été calculés à l'aide de techniques de régression multiple.On a construit quatre tables, une pour la collection totale, et une pour chacune des trois classes, et ceci pour chacun des dix couples d'années. Chaque table contient les valeurs de N(m), désignées par Y, obtenues dans le cadre des quatre modèles linéaires, celui de Morse

(1) $N(m)=\alpha+\beta m$

et des trois modèles modifiés (2) (3) (4). Dans chaque modèle, la variable m est le meilleur prédicteur de Y avec un coefficient de corrélation d'environ 0,8. L'adjonction des deux variables t et s n'augmente que marginalement la valeur du R^2.

Le tableau 5.4 donne pour chaque classe les équations de régression correspondant aux divers modèles et pour chaque régression la valeur du R^2 entre le nombre moyen de transactions Y et

(1) le nombre de circulations de l'année précédente (m)

(2) le nombre de circulations de l'année précédente, et les couples d'années (m,t)

(3) le nombre de circulations de l'année précédente et le nombre total de transactions de la présente année (m,s)

(4) le nombre de circulations de l'année précédente, les couples d'années et le nombre total de transactions de la présente année (m,t,s)

Collection	Regression Equation	R-squared
All	$Y(m) = 0.369 + 0.336\,m$	0.842
	$Y(m,t) = 0.785 + 0.306\,m - 0.0651\,t$	0.882
	$Y(m,s) = -0.275 + 0.320\,m + 0.566\,s$	0.857
	$Y(m,t,s) = 1.25 + 0.306\,m - 0.083\,t - 0.308\,s$	0.882
Q	$Y(m) = 0.425 + 0.312\,m$	0.809
	$Y(m,t) = 0.871 + 0.287\,m - 0.072\,t$	0.867
	$Y(m,s) = -0.687 + 0.290\,m + 0.967\,s$	0.867
	$Y(m,t,s) = 0.047 + 0.286\,m - 0.04\,t + 0.543\,s$	0.872
D	$Y(m) = 0.402 + 0.319\,m$	0.796
	$Y(m,t) = 0.781 + 0.291\,m - 0.059\,t$	0.831
	$Y(m,s) = -0.121 + 0.305\,m + 0.46\,s$	0.806
	$Y(m,t,s) = 1.41 + 0.291\,m - 0.084\,t - 0.419\,s$	0.833
N	$Y(m) = 0.413 + 0.342\,m$	0.740
	$Y(m) = 1.02 + 0.302\,m - 0.098\,t$	0.832
	$Y(m,s) = -0.659 + 0.315\,m + 0.938\,s$	0.782
	$Y(m,t,s) = 1.53 + 0.304\,m - 0.117\,t - 0.339\,s$	0.831

Tableau 5.4 : Equations de régression pour la collection totale et les trois classes [BESH84]

Il faut noter que les coefficients de régression ne varient pas beaucoup entre la collection considérée dans son ensemble et les trois différentes classes. Les valeurs de α sont comprises entre 0.37 et 0,43,les valeurs de β entre 0,31 et 0,34, ce qui correspond aux estimations de Morse.

En conclusion, l'étude de Beshesti et Tague montre que l'adjonction de la variable temps et du nombre total de transactions par an ne modifie pas les résultats de façon suffisamment significative pour justifier leur inclusion dans le modèle. Cependant cette étude a permis de montrer que le paramètre α était dépendant du temps tandis que β présentait des fluctuations plus aléatoires. Ces deux paramètres présentent des fluctuations suffisamment significatives pour ne

plus être ignorées, ce qui pourrait infirmer les hypothèses de base du modèle de Morse

Les modèles proposés par Beshesti et Tague ont cependant soulevé des interrogations chez certains chercheurs:

Tague et Ajiferuke [TAGU 87] ont cherché à déterminer quel modèle, de Morse (1) ou de Beshesti (2), approximait au mieux l'ensemble de données utilisé par Beshesti, constitué des 11 années de transactions à la bibliothèque de l'Université de Saskatchewan. Leurs tests statistiques ne les conduit à aucun résultat significatif qui permettrait de se décider en faveur de l'un ou de l'autre modèle.

Ils notent surtout que le système de Beshesti présente une lacune: il ne figure aucune indication sur les livres qui ne circulent pas . Seules les données relatives aux livres qui ont circulé au moins une fois ont été relevées, et de ce fait, l'effectif de la collection varie d'année en année. Or le modèle de Morse s'applique généralement à une collection constante en nombre.

Burrell[BURR86, p.122] s'étonne que : "*les auteurs (Beshesti&Tague) ne spécifient pas explicitement les techniques de régression utilisées dans l'analyse des données et par conséquent, on doit exprimer des réserves sur leur méthodologie et leurs conclusions. Par exemple, ils notent le comportement particulier des ouvrages qui circulent fréquemment. Ils montrent qu'en ignorant les grandes valeurs de m, on améliore la qualité de l'ajustement linéaire.Cependant ils ne précisent pas si ces valeurs isolées (m>8) sont utilisées ou non dans les calculs des coefficients de régression*".

On peut, au sujet du calcul de ces coefficients, donner un élément de réponse.

4.Techniques de régression multiple

Considérons le modèle linéaire (2) incluant le temps

$$N(m,t)=\alpha+\beta m+ct$$

Bien que Beshesti ne le précise pas, on supposera qu'il n'a pas inclus dans ses calculs les valeurs de N°(m,t) -valeurs observées de la circulation moyenne- pour m>8-, puisque ces valeurs réduiraient l'approximation linéaire (d'ailleurs les valeurs élevées du R^2 obtenues pour ce modèle laissent à penser qu'il a bien fait cette troncature)..Puisque les données portent sur 11 années d'observations, soit dix couples d'années, on dispose donc de 80 valeurs de N°(m,t) (m=1,...8 et t=1,...,10) à partir desquelles on doit calculer les coefficients α,β et c.

Soit b le vecteur colonne contenant les trois coefficients

$$b=\begin{pmatrix}\alpha\\ \beta\\ c\end{pmatrix}$$

Soit Y le vecteur colonne contenant les 80 valeurs théoriques N(m,t)
et A la matrice à 80 lignes et 3 colonnes définis ci-dessous:

$$Y= \begin{bmatrix} N(1,1) \\ N(2,1) \\ N(3,1) \\ N(4,1) \\ N(5,1) \\ N(6,1) \\ N(7,1) \\ N(8,1) \\ N(1,2) \\ N(2,2) \\ \vdots \\ \vdots \\ N(8,2) \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ N(1,10) \\ N(2,10) \\ \vdots \\ \vdots \\ N(8,10) \end{bmatrix} \qquad A= \begin{bmatrix} 1\ 1\ 1 \\ 1\ 2\ 1 \\ 1\ 3\ 1 \\ 1\ 4\ 1 \\ 1\ 5\ 1 \\ 1\ 6\ 1 \\ 1\ 7\ 2 \\ 1\ 8\ 1 \\ 1\ 1\ 2 \\ 1\ 2\ 2 \\ \vdots \\ \vdots \\ 1\ 8\ 2 \\ \vdots \\ \vdots \\ \vdots \\ \vdots \\ 1\ 1\ 10 \\ 1\ 2\ 10 \\ \vdots \\ \vdots \\ 1\ 8\ 10 \end{bmatrix}$$

Le modèle (2) peut alors s'écrire sous forme matricielle::

$$Y=Ab$$

L'estimation du vecteur b des paramètres se fait à l'aide de la formule suivante [TASSI90, p.134]:

$$b= ({}^tA\ A)^{-1}\ ({}^tA)Y°$$

où ${}^tA$ est la transposée de la matrice A et Y° le vecteur colonnes contenant les les 80 valeurs des circulations moyennes N°(m,t) observées.En faisant le produit ${}^tA\ A$, on obtient:

$$ {}^{t}A\,A = \begin{bmatrix} 80 & 360 & 440 \\ 360 & 2040 & 1980 \\ 440 & 1980 & 3080 \end{bmatrix} $$

Il reste à prendre l'inverse de cette dernière matrice et à la multiplier par la matrice transposée de A puis par le vecteur Y°. Comme nous ne disposons pas des valeurs N°(m,t) sur les dix années d'observations, il nous est malheureusement impossible de mener le calcul jusqu'au bout pour obtenir les valeurs des paramètres.

VI. Les distributions géométriques

A. La distribution géométrique de la première année de circulation

Considérons une classe de livres de paramètres α et β, et ayant tous été acheté durant la même année. La distribution de la circulation de la première année dépendra de la perspicacité et du flair du bibliothécaire. Si les livres choisis ont la faveur des lecteurs, nombre d'entre eux circuleront beaucoup durant la première année. Supposons que le choix ait été judicieux. Alors d'après Morse [MORS68, p.101], **la distribution de la circulation de la première année est géométrique de paramètre γ.**

En d'autres termes, la fraction $P_1(\geq m)$des livres qui circuleront m fois ou plus durant leur première année sera

$$P_1(\geq m) = \gamma^m \qquad (6.1)$$

et la fraction des livres qui circuleront exactement m fois sera

$$P_1(m) = P_1(\geq m) - P_1(\geq m+1)$$

$$= (1-\gamma)\,\gamma^m \qquad (6.2)$$

où γ est approximativement égal à R/(1+R), avec R, circulation moyenne durant la première année. Ainsi γ représente la popularité de cette classe particulière (Les points de la courbe notée "1st year" dans la figure 6.1 donnent les valeurs de γ^m pour γ=0,8) . Cependant, si l'allure de la circulation de la première année dépend du jugement et de l'intuition du gestionnaire, l'aspect de la circulation pour les années suivantes dépendra de plus en plus du comportement des usagers, qui est représenté par les valeurs de α et β.

La probabilité $P_2(n)$ pour qu'un de ces livres circule n fois durant la *seconde* année est la somme de la probabilité $P_1(m)$, que le livre ait circulé m fois durant la première année, par la probabilité conditionnelle T_{mn} que le livre circule n fois la seconde année s'il a circulé m fois la première année, et ceci sommé pour toutes les valeurs possibles de m (il s'agit de la formule des probabilités totales).Par le même raisonnement on obtient la probabilité $P_3(m)$

pour la troisième année de circulation, en fonction des circulations de la deuxième année etc...

$$P_2(n) = P_1(0)\, T_{0n} + P_1(1)\, T_{1n} + P_1(2)\, T_{2n} + \ldots$$

$$= (1-\gamma)\,(T_{0n} + \gamma T_{1n} + \gamma^2 T_{2n} + \ldots)$$

$$P_3(n) = P_2(0)\, T_{0n} + P_2(1)\, T_{1n} + P_2(2)\, T_{2n} + \ldots$$

On en déduit que pour une t-iéme année de circulation (t>1), la probabilité pour qu'un des livres de la classe circule n fois sera

$$P_t(n) = P_{t-1}(0) T_{0n} + P_{t-1}(1)\, T_{1n} + P_{t-1}(2)\, T_{2n} + \ldots \qquad (6.3)$$

Et la probabilité que la circulation durant l'année t soit plus grande ou égale à m est la somme

$$P_t(\geq m) = P_t(m) + P_t(m+1) + P_t(m+2) + \ldots \qquad (6.4)$$

où les $P_t(m)$ se calculent à l'aide de l'équation (6.3).

Les points des différentes lignes de la figure 6.1 donnent les valeurs de $P_t(\geq m)$ pour différentes valeurs de m et de t, et pour γ=0,8, α=0,4, et β=0,5 . Ils montrent que la circulation diminue avec le vieillissement des livres.

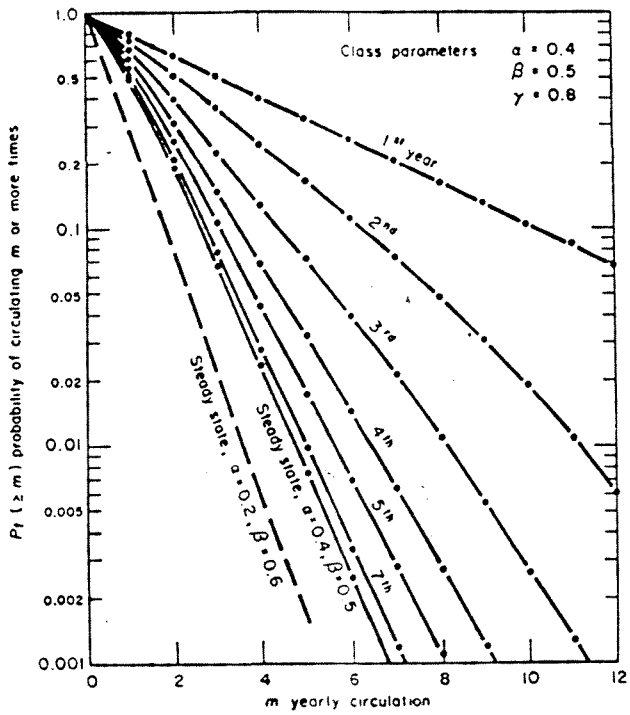


Figure 6.1 : fraction $P_t(\geq m)$ de livres d'une classe qui circuleront m fois ou plus durant l'année t. [MORS68]

On voit que durant la première année 10% de ces livres circulent 10 fois ou plus. A partir de la seconde année, ces circulations ont déjà chuté considérablement; seulement 2% de ces livres circulent 10 fois ou plus pendant la seconde année. Et seul 1 livre sur 400 circulera 10 fois ou plus durant la troisième année.

La circulation décroît fortement avec l'âge des livres, et avec le temps de plus en plus d'ouvrages ne seront empruntés que quelques fois l'an. Cependant, comme nous l'avons expliqué précédemment, il y a toujours une chance pour que certains livres connaissent un regain de popularité; mais cette possibilité diminue elle aussi à mesure que le livre vieillit.

Toutefois, cette baisse de popularité des ouvrages finit par cesser, tant que les paramètres α et β restent constants. Nous constatons sur la figure 6.1 qu'avec les années les courbes de circulations sont de plus en plus rapprochées,pour atteindre finalement une distribution stationnaire notée "Steady state, α=0,4, β=0,5". Au bout d'environ 10 ans la distribution de la circulation demeure la même pour les livres de paramètres α=0,4et β=0,5 ; dans ce groupe de livres, la moitié des ouvrages ne circulera pas du tout ($P_t(\geq 1)$=0,5) ; seul 1 ouvrage sur 5 ($P_t(\geq 2)$ = 20%)circulera deux fois ou plus par an, 1 sur 50 circulera quatre fois ou plus.

B. Le modèle de Morse modifié

Les données de Morse collectées au MIT ainsi que celles rassemblées par Chen à la Countway Library of Medicine semblent suggérer que la fraction des livres qui ne circulent pas - les "no-use"- (c'est-à-dire le cas m=0 dans l'équation $N(m)=\alpha+\beta m$) semble s'éloigner considérablement d'une distribution géométrique. Jain [JAIN67], qui a étudié des groupes homogènes de livres de diverses disciplines, telles que Chimie, Physique et Pharmacie à l'Université de Purdue (1967) a pu lui aussi observer que la classe des "no use" ne suivait pas la même loi de probabilité que les autres classes.

Le modèle de Morse peut en effet s'avérer insuffisant ou biaisé, notamment lorsque la proportion d'ouvrages anciens dans un fonds est importante, ce qui ne manque pas de se traduire par des effectifs élevés au niveau de la classe m=0.

Dans un cas semblable, Morse et Chen [MORS75] proposent de modifier le modèle initial en séparant le fonds actif du fonds inactif. C'est qu'outre sa fâcheuse incidence sur le plan des données statistiques, il est important de connaître et de mesurer le fonds inactif ; accumulant les coûts de stockage mais surtout interférant de façon parfois fort dangereuse avec des ouvrages de plus grande valeur pour les usagers, les ouvrages inactifs peuvent par leur nombre constituer un excellent indicateur de l'inadéquation d'un fonds à la demande et de l'efficacité d'une politique d'acquisition.

1. Analyse des livres actifs

Pour une classe de livres donnée, on observe la circulation des livres actifs (les livres ayant circulé au moins une fois) pendant une période t. On enregistre les nombres N(j) (j≥1),de livres qui ont été empruntés j fois au cours d'une année t (j ≥1).

Morse [MORS68, MORS72] ainsi que Goyal [GOYA70,pp. 20-30] ont montré que **la circulation des livres actifs d'une classe sur une année t donnée suit une distribution géométrique modifiée**. Cela signifie que le nombre de livres qui ont circulé j fois durant l'année t est approximativement égal à Nj(t), avec

$$Nj(t) = Na(t)\,[1-\gamma(t)]\,[\gamma(t)]^{j-1} \qquad (j \geq 1) \qquad (6.5)$$

Cette distribution géométrique modifiée satisfait la condition de normalisation

$$Na(t)=\sum_{j=1}^{\infty} N(j)$$

Les nombres Nj(t) sont les valeurs *estimées* des N(j) . Et N_a est le nombre total exact de livres qui ont circulé au moins une fois pendant l'année.

Une fois qu'une liste des N(j) semblable à celle du tableau 6.1 a été relevée pour la classe de livres étudiée, on peut calculer la valeur de γ pour cette classe en utilisant l'équation (6.6); cette formule est obtenue à l'aide de l'équation (6.5) dans laquelle on prend j=1, avec la valeur exacte de $N_1(t)$, c'est-à-dire N(1); par conséquent on aura l'égalité $N_1(t)=N(1)$.

$$\gamma(t) = 1 - \frac{N(1)}{Na(t)} \qquad (6.6)$$

Le paramètre γ ne change pas beaucoup d'une année sur l'autre pour une classe de livres donnée. Il peut cependant considérablement varier d'une classe à l'autre et d'une bibliothèque à l'autre.

Tableau 6.1: Distribution de la circulation de la classe WM de la Countway Library[1]
[MORS75]

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sum
$N(j)$	583	307	209	152	89	49	32	25	12	4	1	1	1	0	1	1	1,467*
Geom.	583	351	212	129	77	46	28	17	10	6	4	2	1	1	0	0	1,467

*The total number of active books in class WM in Countway Library for this year is therefore N_a = 1,470.

Une fois la valeur de γ calculée, on peut obtenir les diverses caractéristiques de la circulation de la classe pour l'année étudiée en utilisant les équations (6.7) : la circulation moyenne annuelle $R_a(t)$ des livres actifs ; le nombre $N(>m,t)$ des livres de la classe qui ont circulé plus de m fois pendant l'année t ; et la fraction $F(>m,t)$ que représente la circulation de ces livres (qui ont circulé plus de m fois) par rapport à la circulation totale :

$$\bullet \quad R_a(t) = (1/Na(t)) \sum_{j=1}^{\infty} j\, N_j(t) \cong 1/(1-\gamma(t))$$

$$\bullet \quad N(>m, t) = \sum_{j=m+1}^{\infty} N_j(t) \cong N_a(t)\, [\gamma(t)]^m \qquad (6.7)$$

$$\bullet \quad F(>m, t) = 1/(NaRa) \sum_{j=1}^{\infty} j\, N_j(t) = [m(1-\gamma(t)) + 1]\, [\gamma(t)]^m$$

[1]Ces données sont en fait des données non biaisées obtenues en utilisant le facteur de correction de Chen[CHEN76], mais la démarche s'applique aussi bien aux données biaisées

La valeur de γ pourrait aussi être obtenue à partir de l'équation de R_a. Mais comme le précisent Morse et Chen, cette valeur pourrait exagérer l'effet des fluctuations des Nj(t) pour de grandes valeurs de j.

La deuxième ligne du tableau 6.1 donne les valeurs de Nj obtenues avec l'équation (6.5). Une comparaison entre les valeurs estimées des N(j) et les valeurs exactes montre que l'adéquation semble satisfaisante.

La valeur de γ obtenue à partir des données du tableau 6.1 est 1-(583/1 467) = 0,6 026. En conséquence, la circulation moyenne annuelle des 1 467 livres actifs est (1 467/583)=2,516, ce qui est relativement élevé. La circulation totale R_aN_a de tous les livres de la classe WM de la bibliothèque de Countway, entre le 1er Février 1972 et le 31 Janvier 1973 était donc approximativement 2 516, ce qui est assez élevé. En utilisant la deuxième et la troisième équation de (6.7), on voit que $N(>4) = 1\ 467\ (0{,}6026)^4 \cong 190$ livres parmi les 1 467 actifs ont circulé plus de quatre fois pendant l'année, et que ces livres représentent environ (compte tenu de $F(>4) \cong 0{,}34$) un tiers de la circulation totale de la classe.

On a estimé que le nombre total N de livres de la classe WM de Countway s'élevait à environ 2 160 pour l'année 72-73. Par conséquent, la fraction active de cette classe pour l'année en question était d'à peu près 0,68. Cette valeur est assez élevée, mais n'est pas significative en elle-même : à la bibliothèque scientifique du MIT, la fraction active des livres de Physique était seulement de 0,63 en 1962; et c'était la fraction la plus élevée de toutes les classes de la bibliothèque. Cependant, les Na=2 260 livres actifs de physique avaient un taux de circulation moyen Ra=4 plus élevé que ceux de la classe WM. Aussi la circulation totale annuelle de ces 3 700 livres de physique était d'environ 9000, tandis que celle des 2 200 livres de la classe WM était de 3 700.

2. Analyse des livres inactifs

On peut aussi tirer des indications sur les livres restants de chaque classe qui ne circulent pas du tout pendant une année t donnée. On peut, par exemple, estimer le nombre de livres *potentiellement actifs* d'une classe : ce sont les livres qui d'ordinaire circulent , mais qui par hasard n'ont justement pas été empruntés durant l'année d'observation, ce qui est inhérent au caractère aléatoire de la circulation dans le temps.Ces livres potentiellement actifs constituent un sous-groupe des inactifs. Morse et Chen les distinguent des "*remainder* or *remaining* books", qu'on pourrait appeler les laissés-pour-compte : ce sont des livres qui circulent très rarement - ouvrages démodés, anachroniques, ou si pointus qu'ils ne sont empruntés qu'occasionnellement par des spécialistes-.

La circulation des livres potentiellement actifs s'inscrit, avec les livres actifs, dans un simple processus de Markov, comme nous le verrons au paragraphe 3. Dans ce cas, les paramètres de Markov α et β sont calculés sur les livres actifs de la classe. Selon ce modèle, il y a une probabilité non nulle pour qu'un livre qui n'a pas circulé durant une année t circule durant les années suivantes. Inversement, une partie des livres qui ont circulé durant l'année t, ne circulera pas l'année suivante . Autrement dit, **les livres potentiellement actifs d'une année sont susceptibles d'être les livres actifs de l'année suivante.**

De tous les $N_M(t)$ livres actifs et potentiellement actifs de l'année t, une certaine part d'entre eux, $N_p(t)=N_M(t)-N_a(t)$ -les potentiellement actifs-, ne circule pas. Durant l'année t+1, le nombre $N_p(t+1)$ de livres qui ne circuleront pas ne sera pas constitué des mêmes livres qui n'ont pas circulé l'année t, et d'ex-actifs viendront se joindre au groupe des potentiellement actifs; néanmoins le nombre d' inactifs $N_p(t+1)$ ne différera pas beaucoup de l'ancienne valeur $N_p(t)$.

Puisque les données recueillies, de par leur nature, ne nous permettent pas de faire le compte des $N_p(t)=N(0)$ livres potentiellement actifs, on doit utiliser un procédé pour estimer ces valeurs.

On obtient les valeurs approchées des Np en traitant chaque valeur N(j) séparément . Morse et Chen supposent que le nombre N(j) de livres qui circulent exactement j fois pendant l'année t suit un processus de Poisson de moyenne j : on peut alors se demander quelle est la probabilité pour qu'il advienne que des livres qui circulent ordinairement j fois ne circulent pas du tout pendant l'année t ; Cette probabilité est e^{-j} ; alors une estimation grossière du nombre N(j,0) de livres qui ont un potentiel de j circulations mais qui n'ont pas circulé durant l'année t, est reliée à la valeur N(j) des livres qui ont circulé exactement j fois l'année t, par l'équation

$$N(j,0) = N(j)/(e^{j}-1), \qquad N(0) = \sum_{j=1}^{\infty} N(j,0) \cong N_p(t)$$

$$N_M(t) = Na(t)+Np(t) \cong \sum_{j=1}^{\infty}[N(j) +N(j,0)] \qquad (6.8)$$

Les valeurs individuelles N(j,0) ne sont pas utilisées; seule la somme Np(t)=N(0) est récupérée ; .

Cette méthode qui consiste à faire des estimations séparément pour chaque valeur de j diminue la contribution à la somme N(0), des livres qui circulent beaucoup,

comme cela devrait être le cas; en effet les livres qui ont une circulation annuelle plus faible ont de plus grandes chances de ne pas circuler une année ou l'autre que les livres ayant une circulation moyenne élevée.

Dans le tableau 6.2 on peut lire sur la première ligne les valeurs de N(j) du tableau 6.1. La seconde ligne donne les valeurs correspondantes N(j,0), estimées à partir de l'équation (6.8), ainsi que leur somme $N(0) \cong Np(t)$, le nombre de livres potentiellement actifs de la classe. La somme $N_M(t)=N_a(t)+N_p(t) \cong 1\ 870$ est le nombre total estimé des livres actifs et potentiellement actifs de la classe WM. Le rapport C des actifs sur les actifs plus les potentiellement actifs est un paramètre utile pour prédire la circulation future.

Tableau 6.2 : Nombre de livres actifs et potentiellement actifs de la classe WM

[MORS 75]

j	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	Sum
$N(j)$	583	307	209	152	89	49	32	25	12	4	1	1	1	0	1	1	1,467 = N_a
$N(j, 0)$	339	48	11	3	1	0	0	0	0	0	0	0	0	0	0	0	402 = N_p

NOTE.—$N_M = N_a + N_p = 1,870$; $C = N_a/N_M = 0.785$; $R_a = 2.52$; $R_M = CR_a = 1.97$. Estimated total yearly circulation = $R_a N_a = R_M N_M = 3,690$.

3. Prévision de la circulation

La probabilité qu'un ouvrage d'une collection donnée circule n fois pendant l'année t+1, $P_n(t+1)$ est reliée à la probabilité qu'il circule m fois pendant l'année t par la relation

$$P_n(t+1)=N_n(t+1)/N_M = \sum_{m=0}^{\infty} P_m(t)\, T_{mn} \qquad (6.9)$$

où $N_n(t+1)$ est la valeur de N(n) du tableau 6.1 pour l'année t+1, et

$$T_{mn} = \frac{(\alpha+\beta m)^n}{n!}\, e^{-(\alpha+\beta m)}, \qquad \sum_{n=0}^{\infty} T_{mn}=1, \qquad \sum_{n=0}^{\infty} n\, T_{mn}=\alpha+\beta m$$

pour les N_M livres actifs et potentiellement actifs d'une classe.

Si la distribution de la circulation des N_M livres est géométrique, comme c'est le cas dans la plupart des collections, $P_m(t)$ est obtenu à partir des $N_m(t)$ de l'équation (6.5):

$$P_m(t) = \begin{cases} Nm(t)/N_M = C(t)[1-\gamma(t)][\gamma(t)]^{m-1} & (\text{si } m>0) \\ [\,(N_M-Na)/N_M] = 1-C(t) & (\text{si } m=0) \end{cases} \qquad (6.10)$$

où $C(t)=Na(t)/N_M$ est la fraction des livres actifs de la classe pour l'année t. L'équation (6.9) peut être utilisée pour calculer la fraction des livres actifs $C(t+1)=Na(t+1)/N_M$ pour l'année t+1, ainsi que leur circulation moyenne $R_a(t+1)$; ainsi on a

$$R_M(t+1) = C(t+1)R_a(t+1) = \alpha_M + \beta R_M(t) \qquad (6.11)$$

et

$$C(t+1) = 1-P_0(t+1)$$

$$= 1-P_0(t)T_{00}-C(t)(1-\gamma)\sum_{m=1}^{\infty} P_m(t)\,T_{m0} \quad (\text{en utilisant (6.9)})$$

$$= 1-[1-C(t)]\,e^{-\alpha} - C(t)(1-\gamma)\,e^{-\alpha}\sum\gamma^{m-1}\,e^{-m\beta}$$

$$= (1-e^{-\alpha}) + C(t)\,e^{-\alpha}\left[\,1 - \frac{1-\gamma}{(e^{\beta}-\gamma)}\,\right]$$

Finalement

$$C(t+1) = 1-e^{-\alpha}\left[\,1 - \frac{C(t)}{1+J(t)}\,\right] \qquad (6.11)$$

avec $J(t)= \dfrac{1-\gamma(t)}{e^{\beta}-1}$

Le paramètre γ pour les livres actifs et potentiellement actifs de la classe peut être obtenu à l'aide des données du tableau 6.1 et au moyen de l'équation (6.6).
$C(t)=N_a(t)/N_M$

Les valeurs des paramètres α et β pour les livres actifs peuvent être obtenues des données concernant les circulations des années précédentes. L'expérimentation qui a été menée à la bibliothèque scientifique du MIT a montré que si la valeur moyenne de β était à peu près la même pour les livres actifs ou potentiellement actifs d'une même classe, la valeur de α pour les livres potentiellement actifs était petite, proche de zéro. En conséquence, Morse et Chen ont suggéré que les valeurs de α et β qu'il fallait utiliser dans les équations (6.11) et (6.12) étaient les suivantes:

α_M=C(t)*(valeur de α_a obtenue à partir des livres actifs)

β=(valeur de β obtenue à partir des livres actifs)

VII Conclusion

Les modèles de Morse que nous venons d'étudier tiennent compte à la fois de l'obsolescence qui affecte les collections et les phénomènes de regain de popularité qui peuvent toucher certains livres. Ils permettent de prévoir la circulation moyenne à long terme des volumes d'une catégorie; le gestionnaire d'un fonds documentaire bénificie ainsi de précieux indicateurs qui peuvent l'aider dans la mise en oeuvre d'une politique efficace d'acquisition ou de relégation.

Il faut toutefois examiner également les limites de cette technique, si on veut lui attribuer sa vraie place dans le management d'une bibliothèque.

Tout d'abord, la méthode ne tient pas compte de la place disponible dans la bibliothèque ni de la place nécessaire aux ouvrages de différentes catégories, qui est fonction de leur nombre total mais aussi de leur taux de rotation.

Le vieillissement matériel des ouvrages, et la nécessité de les remplacer lorsqu'ils ont été prêtés un certain nombre de fois, ne sont pas pris en compte non plus. Il s'agit pourtant de données aussi indispensables que celles concernant l'obsolescence des ouvrages, surtout si l'on souhaite planifier le nombre d'acquisitions.

Si la méthode permet de déceler un manque d'efficacité des acquisitions, elle ne permet généralement pas de les orienter, car le cadre statistique nécessaire pour l'étude est trop large. Des statistiques plus détaillées sont donc indispensables en complément.

Enfin, la méthode ne tient pas du tout compte du coût des ouvrages : l'unité de base y est le volume. Or ce volume peut être aussi bien un livre de poche qu'un livre d'art

En fait, il semble que ces réserves illustrent la nécessité de combiner plusieurs angles d'approche pour la gestion des bibliothèques. La méthode de Morse constitue un de ces points de vue; elle doit être complétée par d'autres, mais la possibilité qu'elle offre d'une prévision tenant compte des phénomènes de résurgence aussi bien que de l'obsolescence des ouvrages la rend particulièrement intéressante. Et la relative simplicité de sa mise en oeuvre n'est pas un de ses moindres intérêts.

VIII. Bibliographie

BERT90 La gestion dynamique des stocks en bibliothèque : l'analyse de modèles mathématiques.

Bertrand, Annie

Note de synthèse ENSB, Villeurbanne, 1990, 26 p.

BESH84 Morse's model of book use revisited

Beshesti,J and Tague, J.M

Journal of the American Society for Information Science, vol.35, 1984, p.259-267

BURR86 A second note on ageing in a library circulation model : the correlation structure

Burrell, Q.L

Journal of documentation,vol 42, N°2, 1986, p.114-128

CANE87 Le modèle de Morse à la bibliothèque municipale d'Autun

Cane, Simon

Bulletin des Bibliothèques de France, t.32, N°1, Paris, 1987, p.27

CHEN76 Applications of operations research models to libraries

Chen, C C

MIT Press, Cambridge, Mass, 1976

DUCA78 Méthodes du traitement des données bibliométriques pour la gestion des systèmes d'information. Application à l'analyse prévisionnelle de la demande d'ouvrages en bibliothèque

Ducasse, R

Thèse de 3è cycle Sciences de l'information et de la communication. Université de Bordeaux3,1978

DOBR69 The information basis of scientometrics

Dobrov G M, Korennoi A A

A I Michailov et al. (eds), On theorical problems of informatics, Moscow VINITI fo FID, 1969, p.165-191

FUSS61 Patterns in the use of books in large research libraries.

Fussler, H.H and Simon, J.L

Chicago, University of Chicago Press, 1961

GOYA70 Application of operational research to problems of determining appropriate loan periods for periodicals

Goyal, S.K

Libri20, 1970, 94-99

HAWK77 Unconventional use of one-line information retrieval systems : one-line bibliometrics studies

Hawkins DT

Journal of American society for information science, 1977, Vol 28, N°1, p.13-18

JAIN67 A statistical study of book use

Jain, A, K

Ph.D dissertation, Purdue University, 1967

KRAF70 A comment on the Morse-Elston model of probabilistic obsolescence

Kraft, D H

Operations Research, vol.18, 1970, p.1228-1233

MORS68 Library Effectiveness : a system approach

Morse, Philip,M

MIT Press, Cambridge, Mass, 1968, 200p.

MORS72 Measures of library effectiveness

Morse, Philip,M

Library Quaterly, vol.42, 1972, p.15-30

MORS75 Using circulation desk data to obtain unbiased estimates of book use

Morse, P.M and Chen, C.C

Library Quaterly, vol.45, N°2, 1975, p.179-194.

PRIT69 Statistical bibliography or bibliometrics?

Pritchard A

Journal of publication, Vol 25, 1969, p.368-349

RAIS62 Statistical bibliography in the health science

Raisig LM

Bulletin of medical library association, July 1962, Vol 50, N°3, p.450-461

RICH92 Précis de Bibliothéconomie

Richter Brigitte

K G Saur, 1992, 5è éd., 297 p.

ROST93 Veille technologique et bibliométrie : concepts, outils, applications

Rostaing H

Thèse de 3è cycle Sciences de l'information et de la communication, Université de droit et des sciences d'Aix-Marseille, 1993, 353 p.

TAGU87 The Markov end the mixed-Poisson models of library circulation compared

Tague,J and Ajiferuke,I

Journal of documentation, vol. 43, N°3, 1987, p.212-231

TASSI90 Statistique

Tassi, Philippe

Masson, 1990, 278 p.

TRUE64 Two characteristics of circulation

Trueswell, R W

College and Research Libraries, 1964, N°25, p.285-291

SAPO90 Analyse des données et Statistique

Saporta

Technip, 488 p.

ANNEXE A

Generalités sur les processus markoviens

Soit un processus stochastique prenant les valeurs aléatoires $\{X(t_0), X(t_1), \ldots, X(t_n)\}$ en une suite d'épreuves successives. Si à la n-ième épreuve correspond l'aléa x_i, on dira d'un système qu'il est à l'état x_i au temps t_n. Un tel processus aléatoire d'espace d'états E et d'intervalle des temps T est un **processus de Markov d'ordre 1** s'il vérifie l'axiome suivant:

$$P[X(t_n)=x_n / X(t_0)=x_0, \ldots, X(t_{n-1})=x_{n-1}] = P[X(t_n)=x_n / X(t_{n-1})=x_{n-1}]$$

L'axiome de Markov postule que le processus est sans mémoire : la connaissance de l'état du système aux instants consécutifs $t_0, \ldots, t_{n-1}$ antérieurs à tn apporte quant à la connaissance de son état en t_n une certaine information contenue entièrement dans la connaissance de son état le plus récent t_{n-1}.

Un tel processus peut être **homogène** dans le temps si les probabilités de transition ne sont pas affectées par une translation dans le temps. D'autre part, il peut être **discret** : les changements d'états (ou "transitions") ne peuvent intervenir qu'à des isntants donnés, non aléatoires, et au plus en infinité dénombrable.

Un processus de Markov discret $\{X_0, X_1, \ldots, X_n\}$ est appelé Chaîne de Markov dicrète. Dans ce cas l'axiome de Markov s'écrit

$$P[X_n=x_n / X_0=x_0, X_1=x_1, \ldots, X_{n-1}=x_{n-1}]=P[X_n=x_n / X_{n-1}=x_{n-1}]$$

et l'axiome d'homogénéité

$$P[X_t=y / X_s=x]=P[X_{t-s}=y / X_0=x] \qquad \forall (x,y)\in E$$

$$\forall (s,t)\in T$$

Si l'on note $p_{x,y}=P(X_n=y / X_{n-1}=x)$ les probabilités conditionnelles dites de transition, la matrice carrée $P=\{p_{x,y}\}$, où x,y décrivent l'espace d'états E, est appelée matrice de transition. P est une matrice stochastique qui possède les propriétés suivantes

$$\forall\ x,y \qquad p_{x,y} \geq 0$$

$$\forall\ x, \qquad \sum_{y \in E} p_{x,y} = 1$$

On pourra construire la matrice P^n qui distribue les probabilités $p^n_{x,y}$ qu'un système passe d'un état x à un état y en n transitions.

Que se passe-t-il quand n, le nombre de transitions tend vers l'infini? La suite p_n des probabilités de transition tend vers une distribution limite appelée distribution stationnaire. Dans la plupart des cas (il y a des exceptions) la probabilité de transition d'ordre n, $p^n_{x,y}$,devient de moins en moins dépendante de l'état initial lorsque n croît. Cela se traduit par l'égalité des lignes de la matrice P^n