Belgian Journal of Operations Research Statistics and Computer Science

JORBEL

Volume 25 - Number 2-3 - June-September 1985

Revue Belge de recherche opérationnelle, de statistique et d'informatique

Belgisch tijdschrift voor operationeel onderzoek, statistiek en informatica

BELGIAN JOURNAL OF OPERATIONS RESEARCH, STATISTICS AND COMPUTER SCIENCE (JORBEL)

EDITORIAL POLICY

Operations Research, Statistics and Computer Science are of an ever increasing importance as Scientific Methods of Management. They often are linked both in Theory and Practice.

It is the main purpose of Jorbel, the Belgian Journal of Operations Research, Statistics and Computer Science, to promote Research, Applications and Cooperation in these fields.

Therefore the Editorial Policy is to encourage the publication of two kinds of papers: *scientific papers,* such as new theoretical contributions, original applications, computational studies, and on the other hand *didactic papers* such as surveys, tutorials.

Each issue normally includes a tutorial paper. All the papers are refereed. For the sake of effective communication it is recommended to the authors to have their papers written in English. Instructions to the authors are given at the end of this issue.

PUBLICATION

The Belgian Journal of Operations Research, Statistics and Computer Science is published by the Sogesci, rue de la Concorde 53, 1050 Bruxelles - B.V.W.B., Eendrachtstraat 53, 1050 Brussel. It is supported by the "Ministère de l'Education nationale" and the "Ministerie van Nationale Opvoeding".

BELGIAN JOURNAL OF OPERATIONS RESEARCH, STATISTICS AND COMPUTER SCIENCE

REVUE BELGE DE RECHERCHE OPERATIONNELLE, DE STATISTIQUE ET D'INFORMATIQUE

BELGISCH TIJDSCHRIFT VOOR OPERATIONEEL ONDERZOEK, STATISTIEK EN INFORMATICA

Vol. 25 n° 2 et 3 **Vol. 25 nr 2 en 3**

CONTENTS

FOREWORD

For more than 30 years, Professor Jean TECHEM spent most of his time with his students of the "Université Libre de Bruxelles" and the "Faculté des Sciences Agronomiques de l'Etat à Gembloux".

His courses on mathematics, probability and statistics were taught to thousands of students in agriculture, mathematics, physics and psychology. But also, and even mainly, he helped many of them in preparing their final dissertation or doctoral thesis. They all knew that it was never in vain that a student or a young scholar consulted Professor TEGHEM; they knew that his devotion was like his competence.

Some of his former students became collaborators, even colleagues. They are especially grateful for the exceptional instruction and the considerable help they received.

It is these former collaborators and friends who want this issue of the Belgian Journal of Operations Research, Statistics and Computer Science to be published especially in honour of Professor Jean TEGHEM, at the occasion of his retirement and his seventieth birthday.

Fady ABI-KHALIL
Research Associate at the
Centre of Data Analysis and
Stochastic Processes of the
Université Libre de Bruxelles

Pierre DAGNELIE
Professor at the Faculté
des Sciences Agronomiques
de l'Etat à Gembloux

Guy LATOUCHE
Assoc. Professor at the
Université Libre de Bruxelles

Claude LEFEVRE
Assoc. Professor at the
Université Libre de Bruxelles

Jacqueline LORIS-TEGHEM
Professor at the Université
de l'Etat à Mons

Njadi MANYA
Professor at the Université
de Kinshasa

Marc ROUBENS
Professor at the Faculté
Polytechnique de Mons

Raymond SNEYERS
Honorary Departmental
Head at the Royal
Meteorological Institute

Jacques TEGHEM Jr.
Senior Research Associate
at the Faculté
Polytechnique de Mons,
Associate Professor at the
Université Libre de Bruxelles

TRAN-QUOC-TE
Scientific Adviser
at OMNIS

Belgian Journal of Operations Research, Statistics and Computer Science, Vol. 25, n° 2-3.

EXPERIENCE RATING PERTURBED BY A BROWNIAN MOTION

F. ABIKHALIL

CADEPS, CP 135
Université Libre de Bruxelles
50 Avenue F. Roosevelt
1050 - BRUXELLES

ABSTRACT

This paper gives a generalization of a risk process under experience rating in the sense that a Brownian motion is added to the classical model. When the aggregation of claims up to time t, is a diffusion or a compound Poisson process, the probabilities of ruin, both in transient and infinite horizon time, are studied.

1. *INTRODUCTION*

The problem of perturbed experience rating

The principle of experience rating is to adjust premiums continuously (in our paper) on the basis of previous information. Premiums should match the amount of claims and should, at the same time, if possible, take into account the market environment. For example, when the profitability is good, the solvency margin increases to a high level, this stimulates competition and implies new companies drawing up tariff or premium reductions (which suppose that free competition is authorised). Conversely, when the profitability is bad, the insurer should collect more money and consequently increase premiums to face risk exposure. A familiar example is the bonus-malus rating in automobile insurance. For these reasons, we can consider an "experience rating" mathematical model.

Nevertheless, there is a difference between examining premiums in theoretical way and how they actually appear in reality. Actually in practice, the insurer uses "some kind" of experience rating system, which is not based only on risk-theoretical bases but also on other circumstances, let us say, indirect influence factors like :

1) uncertainty on inflation ;
2) Up to date statistics not being available at the time of calculation ;
3) uncertainty due to a lack of precise knowledge about economic activity.

etc.

For example, when industrial and commercial businesses are undergoing a tremendous upswing, this tends to accelerate motor and other traffic, which in turn, tends to increase the number of claims.

On the other hand, during recessions the effects are mainly opposite. So to take into account these indirect influences

we will add, to the "experience rating" model, a perturbation by introducing a Brownian motion for the continuous case considered here (section 3).
In section 4 and 5 we study the case where the aggregate claims up to time t is a Brownian motion with drift and compound Poisson processes respectively.
Moreover, we apply the results of GERBER's paper 1973 [4] to calculate an upper bound for the ruin probability before time t.

Remark
PENTIKAINEN AND RANTALA [9,10] in their studies of the insurance industry in Finland, suggested, in §2.2. Vol.II "Models for premium fluctuation", to perturb the experience rating model with a "white noise" (discrete case) and gave solution to premium calculation for a very simple case.

2. *DESCRIPTION OF THE RISK PROCESS*

We consider a risk process in which the total premiums received in the time-interval $[0,t]$ is denoted by $P(t)$, and $(S(t),\ t \geqslant 0)$ represents aggregation of claims up to time t, we assume that the processes $P(t)$ and $S(t)$ are Markovian and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Finally, let $Z(t)$ be a surplus of a company at time t, $t \geqslant 0$ and write x for $Z(0)$. We have

$$(1) \quad Z(t) = x + P(t) - S(t) \quad , \quad t \geqslant 0.$$

Obviously, $Z(t)$ is one-dimensional Markov process.

3. *EXPERIENCE RATING PERTURBED*

Consider a risk process satisfying (1) except that each element of premium paid is modified by a refund or surcharge according to the stochastic differential equation :

$$(2)\quad dP(t) = (p - k(P(t) - S(t))dt + \sigma dW(t)$$

with $P(0) = 0$ a.s.

and where : (i) p is the base premium constant rate

(ii) $(W(t), t \geqslant 0)$ is a standard Wiener process independent of $(S(t), t \geqslant 0)$

(iii) σ is a positive constant, k being the "experience rating factor" $(0 < k < 1)$.

Equation (2) is a linear stochastic differential equation. From GIHMAN AND SKOROHOD [6] we have the solution

$$(3)\quad P(t) = \exp\left(\int_o^t - k ds\right)\left[\int_o^t \exp\left(-\int_o^t - k du\right).(p + kS(s))ds + \int_o^t \exp\left(-\int_o^t - k du\right)\sigma dW(s)\right]$$

or equivalently

$$(3')\quad P(t) = e^{-kt}\left(\frac{p}{k}(e^{kt} - 1) + k\int_o^t e^{ks}S(s)ds + \sigma X(t)\right)$$

where we define $X(t) = \int_o^t e^{ks}\, dW(s)$.

From the relation (1), it follows that

$$(4)\quad Z(t) = x + \frac{p}{k}(1 - e^{-kt}) + k\, e^{-kt}\int_o^t e^{ks}S(s)\, ds + e^{-kt}\,\sigma X(t) - S(t).$$

In view to characterize and reduce this expression we have the following two propositions.

Proposition 1

X(t) is a gaussian process with zero mean and with covariance :

$$(5)\quad \mathrm{cov}(X(s), X(t)) = \int_0^{\min(s,t)} e^{2ku}\, du$$

For the proof see for example ARNOLD [1] chapter 5.
By elementary computation we can write relation (5) as

$$(5)\quad \mathrm{cov}(X(s), X(t)) = \frac{1}{2k}\,(e^{2k(\min(t,s)} - 1).$$

Let $I(t) = \int_0^t e^{ks}\, d\eta(s)$ where
$(\eta(t),\ t \geq 0)$ is a stochastic process defined on $(\Omega, \mathcal{F}, \mathbb{P})$, and having stationary, independent increments, finite variance with $\eta(0) = 0$ and belonging to $D[0,\infty)$, where $D(0,\infty)$ denote the space of functions on $[0,\infty)$ that are right-continuous and have left hand limites.
We have the following result, justifying the integration by parts for the stochastic integral $I(t)$.

Proposition 2*

The process $(I(t),\ t \geq 0)$ is well defined, a.s. finite, and every sample path satisfies the following relation :

$$(6)\quad I(t) = e^{kt}\eta(t) - k \int_0^t e^{ks}\, \eta(s)ds.$$

Furthermore, $I(t)$ is a.s. in $D[0,\infty)$

---

* This proposition was pointed out by Harrison in [7].

Proof :
Since e^{kt} is a continuous function of bounded variation, we can apply lemma 1 chapter 3 of [2] for the function $\eta(t)$; then the proposition follows form theorem 2 of DUNFORD and SCHWARTZ [3, p.154].

So, from (6) put $\eta(t) \equiv S(t)$ (when S(t) satisfies the conditions on η) we can rewrite (4) as

$$(7)\quad Z(t) = x + \frac{p}{k}(1 - e^{-kt}) - e^{-kt} \int_0^t e^{ks}\, d\,S(s) + e^{-kt}\,\sigma X(t)$$

4. *THE DIFFUSION PROCESS*

Assume now that S(t) satisfies the differential (stochastic) equation :

$$dS(t) = mdt + \sigma_1\, dW_1(t)$$

where m is a constant and $W_1(t)$ is a standard Wiener process independent of W(t).
Then the relation (7) gives

$$(8)\quad Z(t) = x + \frac{p}{k}(1 - e^{-kt}) - e^{-kt} \int_0^t e^{ks}\, mds + e^{-kt}(\sigma_1\, X_1(t) + \sigma X(t))$$

where we define, as before,

$$X_1(t) = \int_0^t e^{ks}\, dW_1(s)$$

From proposition 1 $(X_1(t),\ t \geqslant 0)$, is a gaussian process independent of $(X(t),\ t \geqslant 0)$ with zero mean and as covariance function :

$$\mathrm{cov}(X_1(s),\ X_1(t)) = \frac{1}{2k}\,(e^{2k(\min(t,s))} - 1)$$

It is well known that the sum of two independent gaussian processes is a gaussian one. So we can write :

$$(\sigma_1 X_1(t) + \sigma X(t)) = \hat{\sigma}\hat{X}(t)$$

where (i) $(\hat{X}(t), t \geqslant 0)$ is a gaussian process with zero mean and having the same covariance function of $(X_1(t), t \geqslant 0)$

(ii) $\hat{\sigma}^2 = \sigma_1^2 + \sigma^2$

So write from relation (8)

$$Z(t) = x + \hat{Z}_1(t)$$

with

$$(9)\quad \hat{Z}_1(t) = \frac{p-m}{k}\,(1 - \bar{e}^{kt}) + \hat{\sigma}\bar{e}^{kt}\,\hat{X}(t)$$

It is clear that $(\hat{Z}_1(t), t > 0)$ is a gaussian with independent increments.

4.1. An upper bound on the probability of ruin

We are interested in the variable "time of ruin" defined as usual by :

$$T = \inf \{t \geqslant 0 \ ; \ Z(t) < 0\}$$

Introduce the usual probabilities of ruin, respectively on finite and infinite horizons :

$$\Psi(x,t) = \mathbb{P}\,[\ T \leqslant t \ / \ Z(0) = x],$$
$$\Psi(x) = \mathbb{P}\,[\ T < \infty \ / \ Z(0) = x].$$

GERBER [4] shows that

$$(10)\quad \Psi(x,t) \leqslant \min_r \bar{e}^{rx} \max_{0 \leqslant s \leqslant t} \mathbb{E}[\bar{e}^{r\,\hat{Z}_1(s)}].$$

Now, as

$$\hat{Z}_1(t) \sim \mathcal{N}(m(t),\ s(t))$$

with

$$m(t) = \frac{\mu}{k}\,(1 - e^{-kt})$$

$$s^2(t) = \frac{\hat{\sigma}^2}{2k}\,(1 - e^{-2kt})$$

where $\mu = p - m$,
we can write

$$(11)\quad \mathbb{E}[e^{-r\hat{Z}_1(t)}] = \exp\,[-rm(t) + \frac{1}{2}\,s^2(t)r^2]$$

For fixed t, the exponent in (11) is 0 if $r_1 = r = 0$,

$$\text{or } r = r_2(t) = \frac{4\mu}{\hat{\sigma}^2}\,\frac{1}{1 + e^{-kt}}$$

so, for $r > r_2(t)$ it is positive and increasing.
Consequently, the maximum in (10) is 1 if $0 \leqslant r < r_2(t)$

This reduces (10) to :

$$(12)\quad \Psi(x,t) \leqslant \min_{r \geqslant r_2(t)} \exp\{-rx - r\frac{\mu}{k}(1 - e^{-kt}) + \frac{\hat{\sigma}^2}{4k}(1-e^{-2kt})r^2\}$$

We find (by differentiation) that the minimum is assumed by

$$(13)\quad r_{min} = \frac{2}{\hat{\sigma}^2}\,\frac{kx + \mu(1 - e^{-kt})}{(1-e^{-2kt})}$$

Consequently, we have :

$$(14)\quad \Psi(x,t) \leqslant \exp\{\frac{-1}{\hat{\sigma}^2 k}\,\frac{[kx + \mu(1 - e^{-kt})]}{(1 - e^{-2kt})}\} \text{ if } \frac{\mu(1 - e^{-kt})}{kx} < 1$$

4.2. The ultimate ruin is certain

In order to calculate $\Psi(x)$, recall that

$$(15)\quad Z(t) = e^{-kt}\left[x\, e^{kt} + \frac{\mu}{k}(e^{kt} - 1) + \hat{\sigma}\, \hat{X}_t\right]$$

and define

$$(16)\quad \zeta(t) = x\, e^{kt} + \frac{\mu}{k}(e^{kt} - 1) + \hat{\sigma}\, \hat{X}_t$$

Obviously $T = \inf\{t \geqslant 0 \ ; \ \zeta(t) \leqslant 0\}$.
We have

Proposition 3

$Z(t)$ is a diffusion process with a drift $\mu(y) = \mu^* - ky$, where $\mu^* = \mu + kx$ and an infinitesimal variance : $\sigma^2(y) = \hat{\sigma}^2$

It is clear that $Z(t)$ is gaussian and has continuous sample paths with independent increments, the first two moments of this process are :

$$\mathbb{E}\, Z(t) = x + \frac{\mu}{k}(1 - e^{-kt})$$

$$\mathrm{var}(Z(t)) = \frac{\hat{\sigma}^2}{2k}(1 - e^{-2kt})$$

Thus, we can represent $Z(t)$, by what HARRISON [7] called, compounding Brownian motion,

$$(17)\quad Z(t) = x + \frac{\mu}{k}(1 - e^{-kt}) + \frac{\hat{\sigma}}{\sqrt{2k}}\, W(1 - e^{-2kt}) \qquad t \geqslant 0$$

From this, it follows that Z is a strong Markov with stationnary transition probabilities, so it is a diffusion.
An elementary computation show that

$$\mu(y) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\, \mathbb{E}[Z(t + \Delta t) - Z(t) \ / \ Z(t) = y] = (\mu + kx) - ky.$$

and

$$\sigma^2(y) = \lim_{\Delta t \to 0} \frac{1}{\Delta t}\, \mathbb{E}[(Z(t + \Delta t) - Z(t))^2 \ / \ Z(t) = y] = \hat{\sigma}^2$$

Consequence

In fact $Z(t)$ is an Ornstein-Uhlenbeck (O.U.) process.
To verifie it, let us recall the classical O.U. denoted by $Z_1(t)$:

$$Z_1(t) = e^{-\alpha t} W_1(e^{2\alpha t})$$

$$= e^{-\alpha t} \int_{-\infty}^{t} \sqrt{2\alpha}\, e^{\alpha s}\, dW_2(s)$$

where W_i $i = 1,2$ are two copies of a Brownian motion and α a positif constant.

Obviously, from proposition 3, $Z(t)$ is an O.U. process with $Z(0) = x$ and it is well known that the O.U. process reaches with certainty the exterior of the interval $(0,\infty)$, which implies, for our problem, that the ruin is certain.
Another proof is the following, let us represent

$\zeta(t)$ by the compounding Brownian motion :

$$(19)\quad \zeta(t) = xe^{kt} + \frac{\mu}{k}(e^{+kt} - 1) + \frac{\hat{\sigma}}{\sqrt{2k}} W(e^{2kt} - 1) \qquad t \geqslant 0$$

and let

$$v = e^{2kt} - 1$$

and

$$v^* = e^{2kT} - 1$$

So that v^* is the first $v \geqslant 0$ such that

$$(20)\quad x\sqrt{v+1} + \frac{\mu}{k}(\sqrt{v+1} - 1) + \frac{\hat{\sigma}}{\sqrt{2k}} W(v) = 0 \qquad v \geqslant 0$$

or

$$(21)\quad W(v^*) = -\alpha\sqrt{v^* + 1} + \beta \equiv f(v^*)$$

with

$$\alpha = \frac{\sqrt{2k}}{\hat{\sigma}}\left(x + \frac{\mu}{k}\right)$$

and

$$\beta = \frac{\mu}{k}\frac{2k}{\hat{\sigma}}$$

from the fundamental Wald identity in continuous time, applied to $f(v)$, (see for example SHEPP [12]), it follows that $\Psi(X) = 1$ a.s.

5. *THE COMPOUND POISSON PROCESS*.

Let $S(t)$ be a compound Poisson process; we can write :

$$(22)\quad S(t) = \sum_{i=1}^{N(t)} A_i$$

where $\{A_i\}_{i \geq 1}$ is a sequence of positive independent, identically distrubuted random variables with a common distribution function $F(.)$, and $\{N(t), t \geq 0\}$ is a Poisson stochastic process, independent of the $\{A_i\}_{i\geq 1}$, having parameter λ.

Moreover, we assume $S(t)$ independent of $(X(t), t \geq 0)$, defined in section 3. In the context of classical risk theory :

A_i denotes the amount of the i^{th} claim $(i = 1,2,...)$ and $N(t)$ represents the total number of claims occuring in the time-interval $[0,t]$.

Thus, the Riemann-Stieljes integral $\int_o^t e^{kt}\, d\, S(t)$ becomes :

$$(23)\quad \sum_{i=1}^{N(t)} e^{k\, t_i} A_i$$

where $t_1, t_2, \ldots,$ denote the times at which claims occur.

The surplus process (7) is now :

$$(24)\quad Z(t) = x + \frac{p}{k}(1 - e^{-kt}) - e^{-kt} \sum_{i=1}^{N(t)} e^{kti} A_i + \sigma e^{-kt} X_t$$

or equivalently

$$(25)\quad Z(t) = e^{-kt}[x\, e^{kt} + \frac{p}{k}(e^{kt} - 1) + \sigma X_t - \sum_{i=1}^{N(t)} e^{kti} A_i]$$

$$= e^{-kt}[\tilde{X}_t - X_t^*]$$

where

$$(26)\quad \tilde{X}_t = x\, e^{kt} + \frac{p}{k}(e^{dt} - 1) + \sigma X_t$$

$$(27)\quad X_t^* = \sum_{i=1}^{N(t)} e^{kti} A_i$$

As before, $\tilde{X}_t$ is a gaussian process with independent increments with $\mathbb{E}[\tilde{X}_t] = x\, e^{kt} + \frac{p}{k}(e^{kt} - 1)$

$$\mathrm{var}[\tilde{X}_t] = \frac{\sigma^2}{2k}(e^{2kt} - 1).$$

Consider $\tilde{Z}(t) = Z(t) - x$

Obviously $\tilde{Z}(t)$ is a process with independent increments, then we can apply GERBER's result [4] to calculate an upper bound for $\Psi(x,t)$.
In our case, we have

$$(28)\quad \Psi(X,t) \leqslant \min_r e^{-rx} \max_{0\leqslant s\leqslant t} \exp\, \mathbb{E}\,[e^{-r\tilde{Z}(t)}]$$

Since $(\tilde{X}(t),\ t\geqslant 0)$ and $(X^*(t),\ t\geqslant 0)$ are independent, (28) reduces to :

$$(29)\quad \Psi(x,t) \leqslant \min_{r} e^{-rx} \max_{0\leqslant s\leqslant t} \exp\{-r(x + \frac{p}{k})(e^{ks} - 1) + \frac{\sigma^2}{4k}(e^{2ks} - 1)r^2 + K^*(r,s)\}$$

where $K^*(r,x)$ is the cumulant generating function of X_t^*

From C.G. TAYLOR's paper [13], we have

$$(30)\quad K^*(r,s) = \frac{\lambda}{k} \int_r^{re^{kt}} \frac{\alpha(u) - 1}{u}\, du$$

where $\alpha(u)$ denotes the moment generating function associated with $F(.)$. As in section 4, we can only consider values of r such that $r > r_2(t)$ with $r_2(t)$ being the unique real and positive solution of

$$(31)\quad -r(x + \frac{p}{k})(e^{kt} - 1) + \frac{\sigma^2}{4k}(e^{2kt} - 1)r^2 + K^*(r,t) = 0.$$

Then

$$(32)\quad \Psi(x,t) \leqslant \min_{r\geqslant r_2(t)} \exp[-rx - r(x + \frac{p}{k})(e^{kt} - 1) + \frac{\sigma^2}{4k}(e^{2kt} - 1)r^2 + K^*(r,t)]$$

Example :

if $F(x) = 1 - e^{-x}$ i.e.

negative exponential claim size distribution, then we have

$$(33)\quad K^*(r,t) = \frac{1}{k} \log\{\frac{1 - r}{1 - r\, e^{kt}}\}$$

Some numerical results will be given in the future.

Remarks :

1) When $\sigma = 0$ we have a case treated by Taylor in [13]
2) If we take for $S(t)$ a linear combination of a compound Poisson and Wiener processes (but independent), the whole analysis, in section 4 and 5 is still valid.

REFERENCES

[1] ARNOLD, L., Stochastic Differential Equations (Wiley 1974).

[2] BILLINGSLEY, P., Convergence of Probability Measures (Wiley, New York, 1968)

[3] DUNFORD, N dans SCHWARTZ, J., Linear Operators (1958) (Inter - Science Publishers, New York)

[4] GERBER, H., Martingales in Risk Theory (Mitteilungen der Vereiningung schweizerischer Versicherungsmathematiker 73 205-216).

[5] GERBER, H., The Surplus Process as a Fair Game - Utilitywise (The Astin Bulletin 1974)

[6] GIHMAN and SKOROHOD, Stochastic Differential Equations (Springer Verlag 1972).

[7] HARRISON, J.M., Ruin Problems with compounding Assests (Stochastic Processes and their Applications V.5 n°1 1977).

[8] MEYER, P.A., Probability and Potentials (Blaisdell, Waltham MA, 1966).

[9] PENTIKAINEN, T., Solvency of Insurers and Equalization Reserves, V.I General Aspects (Insurance Publishing Company Ltd, Helsinki).

[10] RANTALA, J., Solvency of Insurers and Equalization Reserves, V.II Risk Theoretical Model (Insurance Publishing Company Ltd, Helsinki).

[11] RUOHONEN, M., On the Probability of Ruin of Risk Processes Approximated by a Diffusion Process (Scand. Actuarial J. 1980, 113-120)

[12] SHEPP, L.A., Explicit Solutions to Some Problems of Optimal Stopping (Annals of Mathematical Statistics 1969, Vol.4

[13] TAYLOR, G.C., Probability of Ruin under Inflationary Conditi or under Experience Raring (The Astin Bulletin 10, 1979).

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

QUELQUES POINTS ESSENTIELS EN CONSULTATION STATISTIQUE

Pierre DAGNELIE

Faculté des Sciences Agronomiques de l'Etat
Statistique et Informatique
Avenue de la Faculté d'Agronomie 8
5800 Gembloux
Belgique

ABSTRACT

This paper stresses some points frequently encountered in statistical consulting, mainly in the field of agricultural and biological sciences.

1. Introduction

La consultation statistique a été une des préoccupations et une des réalisations majeures de la longue et fructueuse carrière du Professeur Jean TEGHEM.

Aussi nous a-t-il paru opportun de consacrer un des articles de cette publication d'hommage au Professeur TEGHEM à quelques points essentiels, qui reviennent à tout moment dans le dialogue entre le chercheur ou l'expérimentateur et le "statisticien-conseil".

Ces quelques points concernent principalement, mais pas exclusivement, la planification des expériences et l'interprétation de leurs résultats, dans le domaine agronomique et biologique. Les principes qui sont développés s'appliquent cependant aussi à d'autres secteurs d'utilisation des méthodes statistiques.

2. Le but et les conditions de l'expérience

La définition précise du but poursuivi et des conditions de travail constitue indiscutablement un des points essentiels de tout échange de vues entre chercheur et statisticien.

A ce stade du dialogue, le statisticien devra s'efforcer d'avoir une vision d'ensemble du ou des problèmes étudiés, depuis les intentions initiales du chercheur jusqu'aux différentes possibilités d'analyse statistique et d'interprétation des résultats, en prévoyant autant que possible les diverses difficultés qui pourront être rencontrées, les solutions qui pourraient être apportées à ces difficultés, etc. (choix des unités expérimentales, définition des mesures à réaliser, etc.).

On notera en particulier combien le fait de prévoir dès le départ l'analyse statistique éventuelle des résultats, en termes d'analyse de la variance et de contrastes par exemple, peut aider le statisticien et son interlocuteur à mieux préciser le ou les objectifs qui seront poursuivis, à mieux identifier les différentes sources de variation auxquelles ils seront confrontés, etc.

3. L'échantillonnage à deux ou plusieurs degrés

L'échantillonnage à deux ou plusieurs degrés est de pratique extrêmement courante, en expérimentation comme dans la réalisation d'enquêtes. Il est aisé d'en citer quelques exemples : choix d'un certain nombre d'arbres dans un verger et prélèvement d'un certain nombre de fruits sur chacun des arbres, choix d'un certain nombre de champs de betteraves dans une région donnée et prélèvement d'un certain nombre de betteraves dans chacun des champs, prélèvement d'échantillons de terre dans une parcelle d'expérience et réalisation de plusieurs analyses au laboratoire sur des sous-échantillons, etc.

Dans de telles situations, la détermination du nombre d'observations à réaliser à chacun des deux ou des différents niveaux de l'échantillonnage est, très souvent, un sujet de contestation entre le statisticien et le chercheur ou l'expérimentateur.

Le point de vue du statisticien sera pratiquement toujours de réduire au strict minimum (2 par exemple) le nombre d'observations relatif aux niveaux inférieurs de l'échantillonnage (nombre de fruits par arbre, nombre de betteraves par champ, nombre d'analyses par échantillon de terre, etc.), en augmentant au maximum le nombre d'unités prélevées aux niveaux supérieurs (nombre d'arbres, nombre de champs, nombre d'échantillons de terre, etc.). Le point de vue du chercheur sera souvent divergent et il appartiendra au statisticien de convaincre au mieux son interlocuteur, plus par des exemples numériques chiffrés qu'à l'aide de formules. Le cas échéant, il y aura lieu de tenir compte aussi des facteurs coûts ou temps [DAGNELIE, 1979-1980].

4. La répétition des expériences dans l'espace et dans le temps

Un autre sujet de préoccupation du statisticien consultant sera la répétition des expériences dans l'espace, en différents lieux, et dans le temps, au cours de différentes années ou périodes de culture. En vue d'aboutir à des conclusions (conseils de fumures par exemple) qui puissent être transposées dans la

pratique, il importe en effet que ces conclusions soient préalablement validées dans des conditions aussi larges que possible.

On notera que cette question, dont l'importance est évidente en matière agronomique, doit également retenir l'attention dans de nombreux autres domaines, pour lesquels la variabilité dans l'espace ou dans le temps dépasse largement la variabilité locale et instantanée. Tel est le cas notamment pour toutes les analyses de laboratoire, dont les résultats présentent le plus souvent des fluctuations inter-laboratoires très supérieures aux fluctuations intra-laboratoires.

Le problème de la répétition des expériences en différents endroits et au cours de différentes périodes se pose dans des termes semblables au problème de l'échantillonnage à deux ou plusieurs degrés. Le but du statisticien sera généralement ici de convaincre son interlocuteur de l'intérêt qu'il y a à augmenter dans toute la mesure du possible le nombre de répétitions dans l'espace et dans le temps, aux dépens du nombre de répétitions réalisées au cours de chacune des expériences [DAGNELIE, 1981].

5. Les limites de l'expérimentation

L'addition de la variabilité dans l'espace et dans le temps, à la variabilité locale et instantanée du matériel expérimental, conduit à des limitations très strictes, dont on prend trop rarement conscience. On peut démontrer par exemple que, dans des conditions particulièrement favorables, où les différentes sources de variation sont de l'ordre de quelques pour cent seulement, les différences de moyennes qu'on peut espérer mettre en évidence sont, malgré toutes les répétitions possibles, de l'ordre de 10 % ou plus, cet ordre de grandeur devant être multiplié au moins par 1,5 ou 2 dans des conditions moins favorables [DAGNELIE, 1981].

Il est donc illusoire le plus souvent, dans le domaine biologique, d'organiser des expériences dont l'objectif serait de mettre en évidence des différen-

ces, de croissance ou de rendement par exemple, de quelques pour cent. Des limites analogues existent certainement, mais à d'autres niveaux sans doute, dans d'autres domaines.

En fonction de ces limites, le meilleur conseil que pourra donner le statisticien sera, dans certains cas, non pas de choisir tel ou tel dispositif expérimental, mais bien de renoncer à toute expérimentation, si les moyens disponibles ne donnent pas des chances raisonnables d'aboutir à des résultats intéressants.

6. L'utilisation de dispositifs expérimentaux simples

Les ouvrages classiques d'expérimentation, et plus encore les revues, regorgent de dispositifs sophistiqués, dont le but principal est de maîtriser au mieux la variabilité du matériel étudié. Il faut noter toutefois que, d'une façon générale, cette maîtrise de la variabilité est purement locale et instantanée, et qu'elle n'a aucune influence, ou pratiquement aucune influence, sur la variabilité dans l'espace et dans le temps.

Dans l'optique évoquée ci-dessus, d'expériences répétées en différents endroits et au cours de différentes périodes, l'utilisation de dispositifs relativement complexes se justifie donc peu. Les dispositifs les plus simples (dispositifs en blocs aléatoires complets par exemple) sont alors, le plus souvent, les plus adéquats, et cela aussi parce qu'ils sont les plus faciles à implanter et les plus robustes.

7. La maîtrise des conditions expérimentales

L'expérimentateur donne fréquemment au statisticien l'impression que la variabilité des résultats qu'il attend est d'autant plus réduite qu'il maîtrise mieux les conditions de son expérience, et notamment que la variabilité des expériences en serres, en chambres de culture ou en laboratoires est plus faible que la variabilité observée dans des conditions naturelles. Mais il faut savoir

que les résultats chiffrés démentent souvent cette impression. Il apparaît en effet que les conditions artificielles des serres, des chambres de culture, etc. sont plus "limites" et que, dans ces conditions, des fluctuations même réduites d'éclairement, de température, d'humidité, etc. peuvent influencer de façon considérable la croissance et le développement des organismes vivants.

Des réflexions analogues peuvent être formulées également en ce qui concerne les méthodes modernes d'analyse chimique. Ces méthodes, de plus en plus élaborées, sont elles aussi de plus en plus sensibles, dans certains cas, à des fluctuations des conditions ambiantes.

Une bonne maîtrise apparente des conditions expérimentales ne dispense donc pas d'une grande prudence dans la planification des expériences et dans l'interprétation de leurs résultats.

8. Les contacts entre l'expérimentateur et le statisticien, et le suivi des expériences

Traditionnellement, les contacts et les échanges de vues entre le statisticien et le chercheur ou l'expérimentateur ont généralement lieu dans le bureau du statisticien, ou même par téléphone seulement. Si cette solution est évidemment la plus confortable pour le statisticien, elle n'en est pas pour autant la plus sûre.

Il n'est pas certain, en effet, que les informations qui sont communiquées dans de telles conditions par l'expérimentateur au statisticien soient suffisamment complètes, ni qu'elles soient bien comprises par ce dernier. De même, il n'est pas certain, loin de là, que les conseils qui sont donnés par le statisticien dans de telles conditions soient les plus judicieux, ni qu'ils soient suffisamment explicites et bien compris par l'expérimentateur, ni a fortiori qu'ils seront bien appliqués par celui-ci ou par ses collaborateurs.

Rien ne vaut, chaque fois que cela s'avère possible, un contact "sur le terrain" (au champ, dans la serre, au laboratoire, etc.), et cela non seulemen

pour l'expérimentateur, surtout débutant, mais aussi pour la formation personnelle du statisticien.

Les contacts "sur le terrain" pourront d'ailleurs avantageusement se poursuivre au-delà de la planification des expériences, en vue d'assurer au mieux le suivi de celles-ci.

9. Les situations imprévues

Le suivi des expériences conduit, plus fréquemment qu'on ne le croit, à observer des anomalies par rapport aux principes théoriques, notamment en ce qui concerne la répartition aléatoire des objets. Ces anomalies peuvent être soit délibérées, et parfois justifiées, soit tout à fait accidentelles.

Il importera de toujours tenir compte de ces situations anormales au moment de l'analyse des résultats ou, au moins, de s'assurer du fait que ces situations sont sans conséquences en ce qui concerne l'interprétation des résultats.

Les méthodes de simulation peuvent être fort utiles à cet égard [CLAUSTRIAUX, 1981a, 1981b] et on notera également, à ce propos, l'intérêt du chapitre "the spoilt experiment" figurant dans le dernier livre de PEARCE [1983].

10. L'autopsie des expériences

Très souvent, l'analyse des résultats d'expériences s'arrête au constat de l'absence ou de l'existence de différences significatives entre moyennes ou entre pourcentages. Parfois, viennent cependant s'ajouter à cela quelques calculs de limites de confiance ou de courbes ou surfaces de réponse.

Mais les moyens modernes de traitement de l'information permettent d'aller beaucoup plus loin, dans le sens de l'estimation de composantes de variances, du calcul de résidus, de l'examen de ces résidus, de l'étude de leur distribution, etc. Tous ces éléments permettent d'effectuer a posteriori une analyse critique de l'expérience, qui peut s'apparenter à une véritable autopsie [CLAUSTRIAUX, 1983; PEARCE, 1976].

Si cette opération ne présente pas toujours un intérêt immédiat important elle s'avère le plus souvent très utile en vue de la planification éventuelle d'expériences ultérieures. En particulier, la connaissance d'ordres de grandeu de composantes de variances est un élément de base de tout dialogue entre expérimentateur et statisticien.

11. L'utilisation de l'ordinateur

Bien qu'en fait, il ne s'impose pas toujours, l'emploi de l'ordinateur en vue de l'analyse des résultats d'expériences est devenu une pratique courante. Mais cet emploi n'est certainement pas sans danger.

Un premier danger concerne l'enregistrement des données sur support informatique (perforation, encodage, etc.), et en particulier la collecte automatique des données ("data capture"). Certaines procédures d'enregistrement ou de saisie des données n'offrent, en fait, que peu de possibilités de contrôle de leur validité. Il appartiendra alors au statisticien, comme à l'expérimentateur, d'être particulièrement circonspect dans l'interprétation des résultats fournis par l'ordinateur. L'examen des résidus constitue, à ce stade de l'analyse, un outil particulièrement utile.

Un deuxième danger est lié à l'utilisation de logiciels "tout faits". Cette utilisation peut en effet conduire à l'emploi de méthodes d'analyse statistique inadéquates, qui ne correspondent pas à l'objectif défini initialement ou au dispositif expérimental choisi. D'autre part, le recours à certains logiciels "clef en main" peut aussi conduire à l'utilisation de méthodes statistiques, peut-être bien adaptées, mais inconnues ou mal connues de l'expérimentateur, avec en conséquence un risque non négligeable de mauvaise interprétation ou d'interprétation abusive des résultats.

12. En guise de conclusion

Nous avons ainsi parcouru, très rapidement, l'ensemble de la boucle qui va de la définition du but et des conditions d'une expérience ou d'un groupe d'expériences à l'obtention des résultats, et même presque à la planification de l'expérience ou des expériences suivantes.

Ce rapide tour d'horizon permet de souligner l'importance d'un contact régulier entre le statisticien et l'expérimentateur et, aussi, mais ceci est un sentiment plus personnel, le très grand intérêt que présente pour le statisticien le travail de consultation.

A l'intention du lecteur qui souhaiterait être documenté plus complètement, nous ajoutons ci-dessous, aux références bibliographiques citées dans le texte, quelques références complémentaires [CANTILLON et DAGNELIE, 1979; FINNEY, 1978, 1981; FINNEY et YATES, 1981].

Bibliographie

CANTILLON P. et DAGNELIE P. [1979]. Semaine d'Etude Internationale Statistique et Informatique en Agronomie / International Study Week Statistics and Computer Science in Agriculture. Gembloux, Faculté des Sciences Agronomiques et Centre de Recherches Agronomiques, 238 p.

CLAUSTRIAUX J.J. [1981a]. Influence de différentes randomisations restreintes sur l'analyse des résultats d'expériences agronomiques : principes. Biom. Praxim. 21, 13-27.

CLAUSTRIAUX J.J. [1981b]. Influence de différentes randomisations restreintes sur l'analyse des résultats d'expériences agronomiques : applications. Biom. Praxim. 21, 71-90.

CLAUSTRIAUX J.J. [1983]. Autopsie d'une expérience. Gembloux, Faculté des Sciences Agronomiques, 10 p.

DAGNELIE P. [1979-1980]. Théorie et méthodes statistiques : applications agronomiques (2 vol.). Gembloux, Presses Agronomiques, 378 + 463 p.

DAGNELIE P. [1981]. Principes d'expérimentation. Gembloux, Presses Agronomiques, 182 p.

FINNEY D.J. [1978]. Statistics and statisticians in agricultural research. J. Agric. Sci. 91, 653-659.

FINNEY D.J. [1981]. The misuse of mathematicians, statisticians and computers in agricultural research. Exper. Agric. 17, 345-353.

FINNEY D.J. et YATES F. [1981]. Statistics and computing in agricultural research. In : COOKE G.W. et al. Agricultural research 1931-1981 : a history of the Agricultural Research Council and a review of developments in agricultural science during the last 50 years. London, Agricultural Research Council, 219-236.

PEARCE S.C. [1976]. The conduct of "post-mortem" on concluded field trials. Exper. Agric. 12, 151-162.

PEARCE S.C. [1983]. The agricultural field experiment. New York, Wiley, 335 p.

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

BLOCKING IN TANDEM QUEUES

Brigitte NICOLAS and Guy LATOUCHE

Université Libre de Bruxelles
Séminaire de Théorie des Probabilités, CP 212
Boulevard du Triomphe
1050 Bruxelles
Belgium

ABSTRACT

We consider a system of two queues in tandem with a finite intermediary buffer. We examine the influence of variability in service requirement at the second server, on the behaviour of the system.

Introduction

The queueing model considered here consists of two units in series with a finite intermediary buffer. Arriving customers enter an infinite buffer in front of Unit I. After being served at Unit I, a customer enters a finite buffer in front of Unit II if the buffer is not full; if the buffer is full, the customer under consideration may not leave Unit I, which thereby becomes blocked and unable to process waiting customer. At a later time, Unit I becomes available again, in a manner to be described later.

In many practical situations, e.g. in data communication networks, the use of an intermediary buffer is dictated by the physical necessity of decoupling the functioning of Units I and II. In other circumstances, it may be advantageous to use an intermediary buffer, in order to render each unit less dependent on random fluctuations in the functioning of the other.

Our purpose in the present note is to examine how the variability of service requirements at Unit II influences the functioning of Unit I. The mathematical model and method of analysis are described in the next section. In Section 2 are defined the parameter values chosen for the numerical analysis. Some results are presented and discussed in Section 3.

For a survey of the literature on such systems, we refer to the bibliography in Latouche and Neuts [1], where a similar system is studied.

1. *The mathematical model*

We assume that customers arrive in the system according to a Poisson process with parameter λ; the duration of service at Unit I is exponential with parameter μ; the service at Unit II is phase-type (PH) with representation $(\underline{\alpha},T)$; all random variables are independent.

PH distributions form a general class, defined and extensively analysed in Neuts [2]. In short, a random variable has a PH distribution if it may be represented as the time until absorption for a Markov process with one absorbing state. They are characterized by the number m of transient states

(or phases), the stochastic m-vector $\underline{\alpha}$ that gives the initial probability distribution on the transient states, and the infinitesimal generator T, of order m, that determines transitions among the transient states. Erlang, hyperexponential and Coxian distributions with real, positive parameters, all are special cases of PH distributions.

Customers who have not yet been served at Unit I are called 1-customers; customers who have been served at Unit I but not at Unit II are called 2-customers. The intermediary buffer is finite and we denote by M the maximum number of 2-customers : there are at most M-2 such customers in the buffer, one being served at Unit II, and one unable to leave Unit I when the buffer is full.

We assume that when Unit I becomes blocked, it stays so until there remain K 2-customers in the system, $0 \leqslant K \leqslant M-1$. The case when K=M-1 means that Unit I operates again as soon as one 2-customer finishes its service, thereby releasing one space in the buffer and allowing the blocking customer to leave Unit I. The case when K=0 means that the buffer and Unit II must become completely empty before Unit I may start functioning again.

The quantity M-2 may be thought of as a technological constraint on the buffer size, while K determines a control policy, to be used e.g. if there are costs associated to the shutting off and starting up of Unit I.

Under the stated assumptions, the system may be described as a Markov process on the state space $\{(n,i,j); n \geqslant 0,\ i=0,1,\ldots,M-1,M',(M-1)',\ldots,(K+1)';\ 1 \leqslant j \leqslant m\}$, where n is the number of 1-customers, i is indicative of the number of 2-customers and the state of Unit I (a symbol ' meaning that Unit I is blocked), and j is the service phase at Unit II. Since n may change by one unit at most, that Markov process is a quasy-birth-and death process of the type extensively studied by Neuts [2]. The corresponding analysis is well documented in the literature and shall not be reproduced here, for lack of space. The interested reader will find the general theorems in [2].
In Nicolas [3], the theory has been applied to the model at hand, and several specific results have been obtained. To obtain the stationary probability distribution, it is necessary to compute a matrix of order N=(2M-K); the algorithm developed in [3] is such that no other matrix of that order need be stored.

2. *The numerical analysis*

The stochastic process is specified by the following parameters :

- the input rate λ,
- the service rate μ at Unit I;
- the maximum number M of 2-customers;
- the control parameter K;
- the representation $(\underline{\alpha},T)$ of the service distribution at Unit II.

Our purpose is to measure how variability in service requirements at Unit II influences the bahaviour of the system. A frequently used, global measure of variability is the ratio C of the standard deviation to the expected value. We have constructed, for a number of values of C, several PH-distributions with the same value for C. For each, the expected value equals one, thereby the unit of time is fixed.

The queueing system is stable if and only if $\lambda < \lambda_{max}$, where λ_{max} is a non explicit, but easily computed, function of all the other parameters. This we have firstly examined.

We then have studied, for certain values of λ, the queue in front of Unit I and Unit II and both the stationary probability π_o that Unit I is blocked and the stationary probability π_1 that Unit I becomes blocked at the end of a service : π_o and π_1 give different information since the former is a time-average, while the later is a customer-average.

3. *Numerical results*

3.1 The maximal arrival rate. We have systematically observed that it is an increasing function of K, for fixed μ, M and $(\underline{\alpha},T)$, therefore it is best to set K=M-1 in order to maximize the throughput of the system. Also, it is a monotonically increasing function of μ, for fixed K, M and $(\underline{\alpha},T)$: see Figures 1 and 2 where different PH-distributions are identified by their coefficient of variability C. Not surprising, λ_{max} is bounded aboved by the minimum of μ and $(E\ [\text{service at Unit II}])^{-1}$.

3.2 *The stationary distribution of the system* is similarly affected by the variability of the PH-distribution $(\underline{\alpha},T)$. We display on Figures 3,4 and 5 respectively values of m_1, m_2 and π_o, where m_i=E (number of i-customers), i=1,2, and π_o is the stationary probability that Unit I is blocked.

To compare different systems under "equal load" conditions, one may either impose the same rate of arrival λ, or the same ratio $\rho=\lambda/\lambda_{max}$. Because of the large differences in λ_{max} for different distributions $(\underline{\alpha},T)$, one may not confuse the two definitions. Since λ is the real systems parameter, we display m_1, m_2 and π_o as functions of λ. In order to keep part of the information that might be contained in ρ, we have marked each curve by dots corresponding to the values ρ=.3,.5,.7 and .9.

We observe on Figure 4 one instance when it is necessary to properly define the notion of equal load. For a given value of λ, m_2 clearly increases with the variability of the PH-distribution. For fixed ρ=0.9 however, the values of m_2 for each distribution (highest point on each curve) are nearly equal.

We must emphasize that the behaviour of the system depends on the whole distribution $(\underline{\alpha},T)$, and not on the coefficient C only. Results not reproduced here indicate the existence of distributions $(\underline{\alpha},T_1)$ and $(\underline{\alpha},T_2)$ such that $C_{(1)} > C_{(2)}$ but $m_{1(1)} < m_{1(2)}$, $m_{2(1)} < m_{2(2)}$ and $\pi_{o(1)} < \pi_{o(2)}$ for whole ranges of values of λ.

3.3 *The blocking phenomenon* may be measured either by the stationary probability π_o that Unit I is blocked at time t, or the stationary probability π_1 that Unit I becomes blocked after serving a 1-customer. The former measures the length of time spent in the blocked state, the latter measures the frequency of switching from a state where Unit I is available to the state where it is blocked.

The variability in ($\underline{\alpha}$,T) influences differently π_o and π_1 (see Figures 5 and 6). For instance, consider the values of π_o and π_1, for C=2.5 and 5 respectively, and λ=5. We observe that the queue with highest variability is blocked during longer periods of time ($\pi_o(C=5) > \pi_o(C=2.5)$) but changes less frequently from being available to being blocked ($\pi_1(C=5) < \pi_1(C=2.5)$).

This may be interpreted as follows. For the distribution with C=5, services at Unit II are typically very short, with an occasional very long one. When a very long service occurs, Unit I is likely to become blocked and to remain so for a long time. When that long service terminates, Unit II will process many customers with a very short service, during which time Unit I is unlikely to become blocked again.

REFERENCES

[1] LATOUCHE, G. and NEUTS, M.F. : "Efficient algorithmic solutions to exponential tandem queues with blocking", SIAM J. Alg. Disc. Methods 1, 1980, 93-106.

[2] NEUTS, M.F. : "Matrix-Geometric Solutions in Stochastic Models. An algorithmic Approach", The Johns Hopkins University Press, Baltimore, 1981

[3] NICOLAS, B. : "Analyse quantitative d'un système de deux files en tandem avec blocage", Mémoire de licence en mathématiques, Université Libre de Bruxelles, 1981

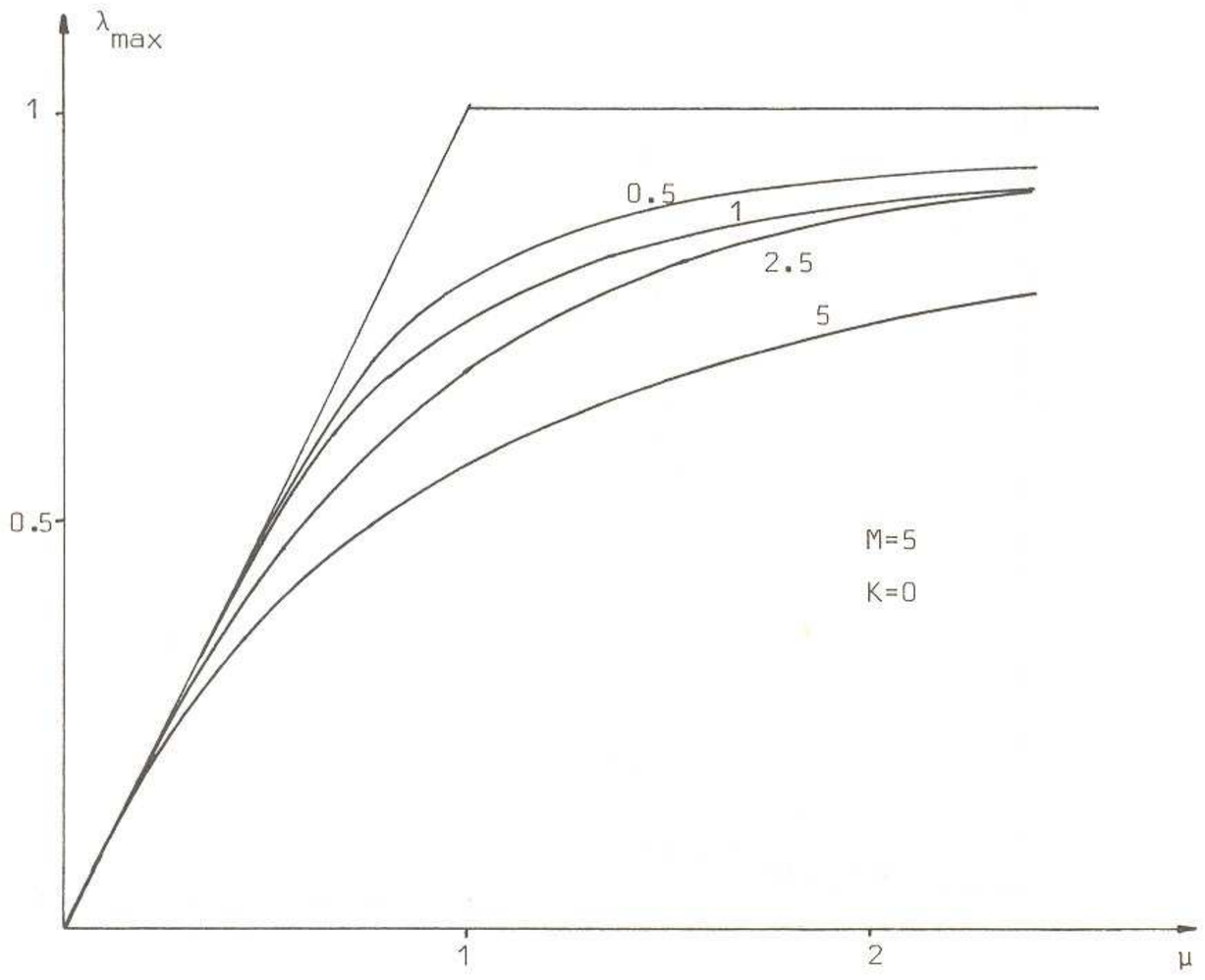


Figure 1. Maximum throughput, K=0

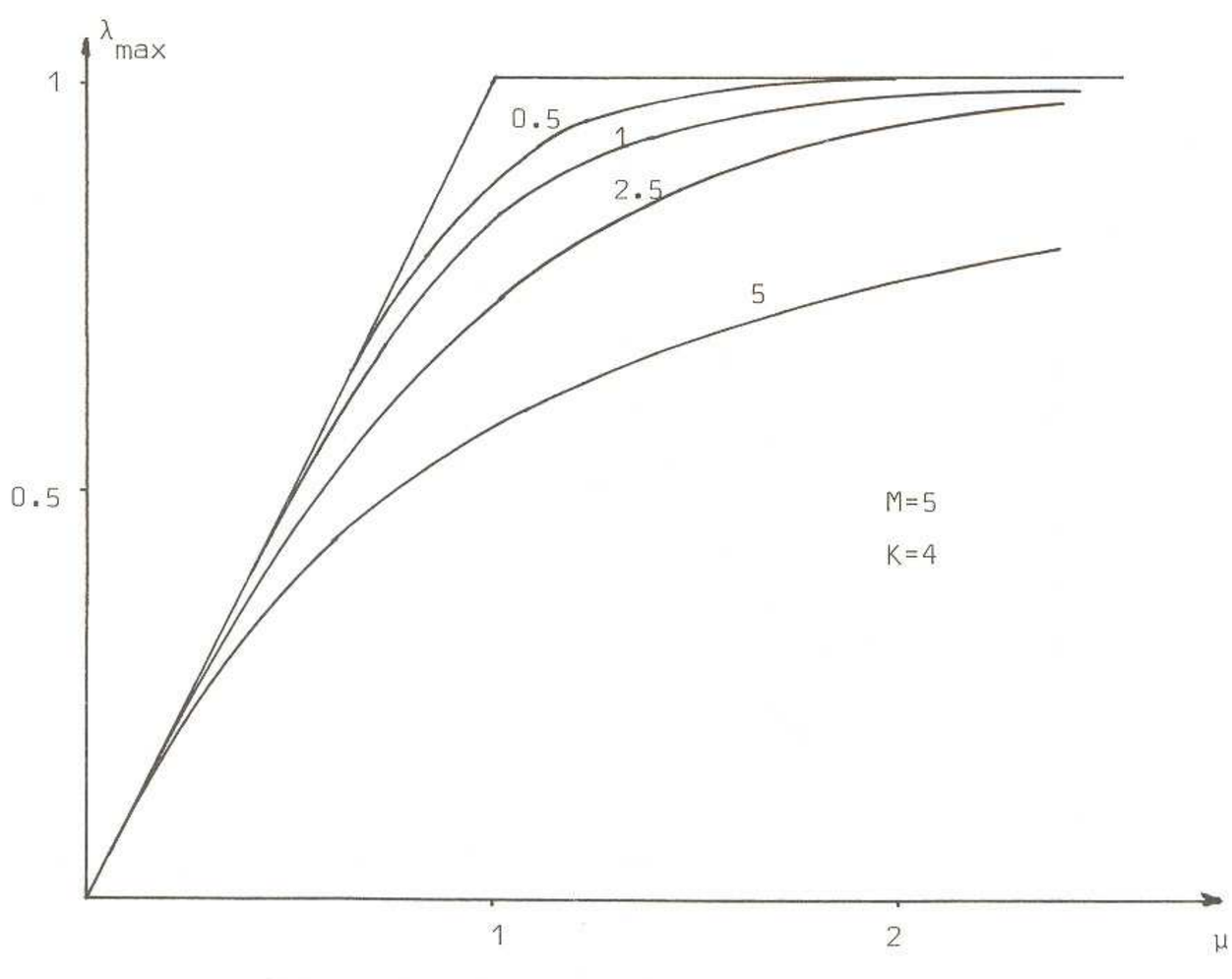


Figure 2. Maximum throughput, K=M-1

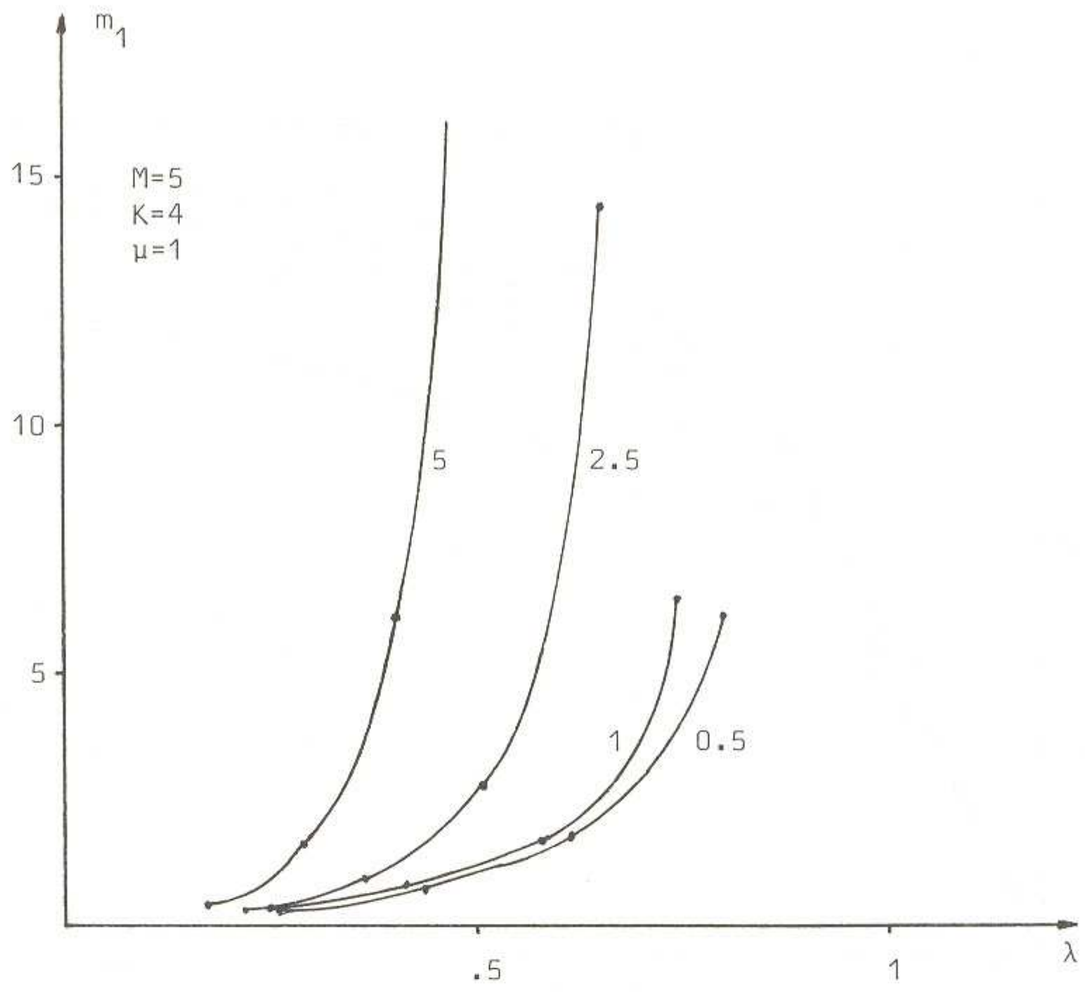


Figure 3. Expected number of 1-customers

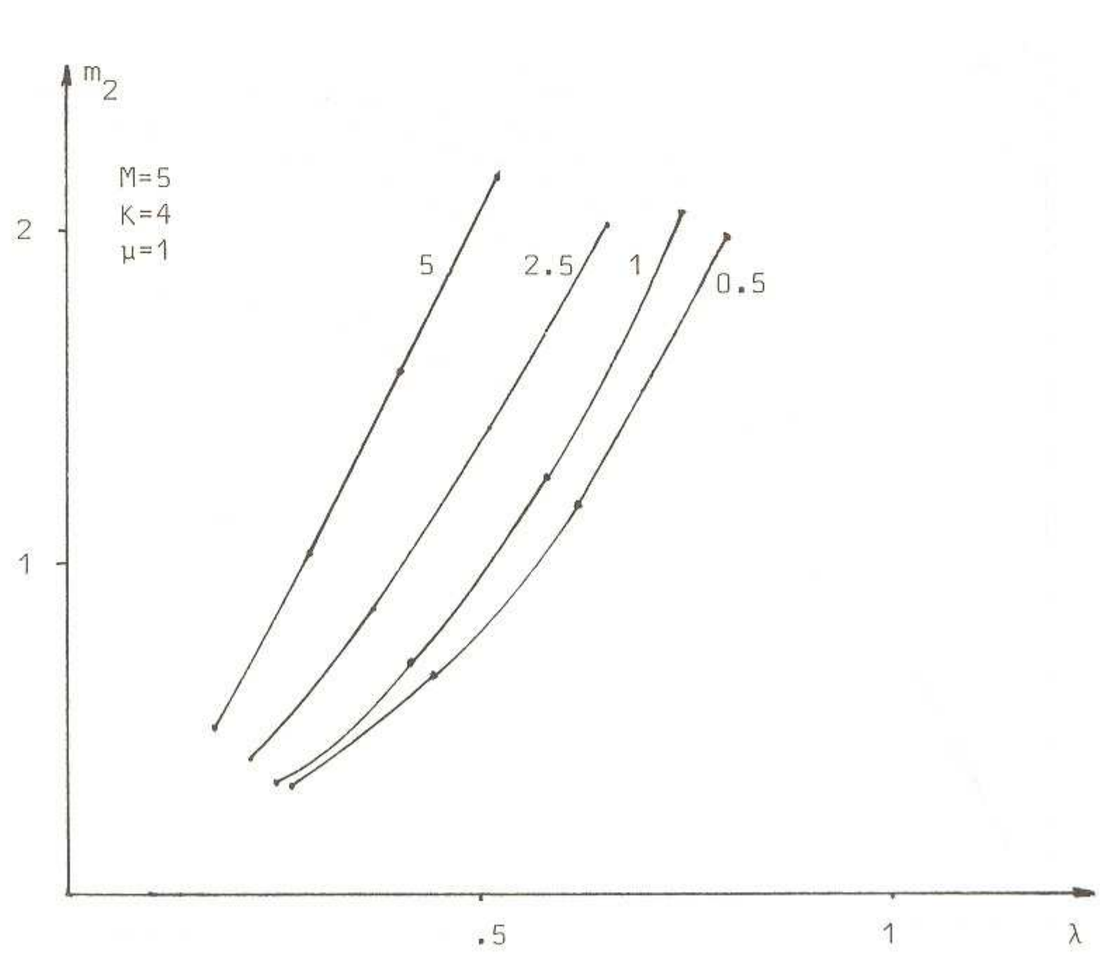


Figure 4. Expected number of 2-customers

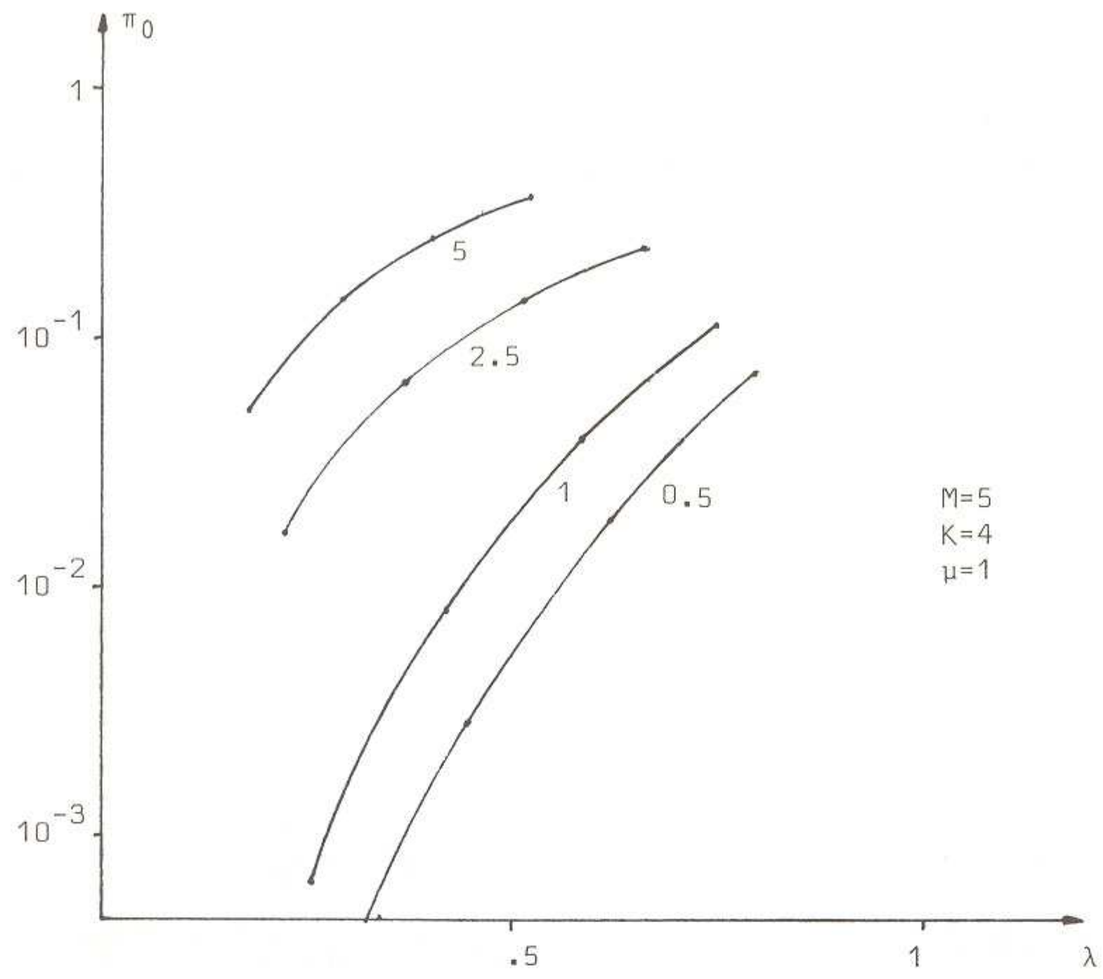


Figure 5. Probability that Unit I is blocked

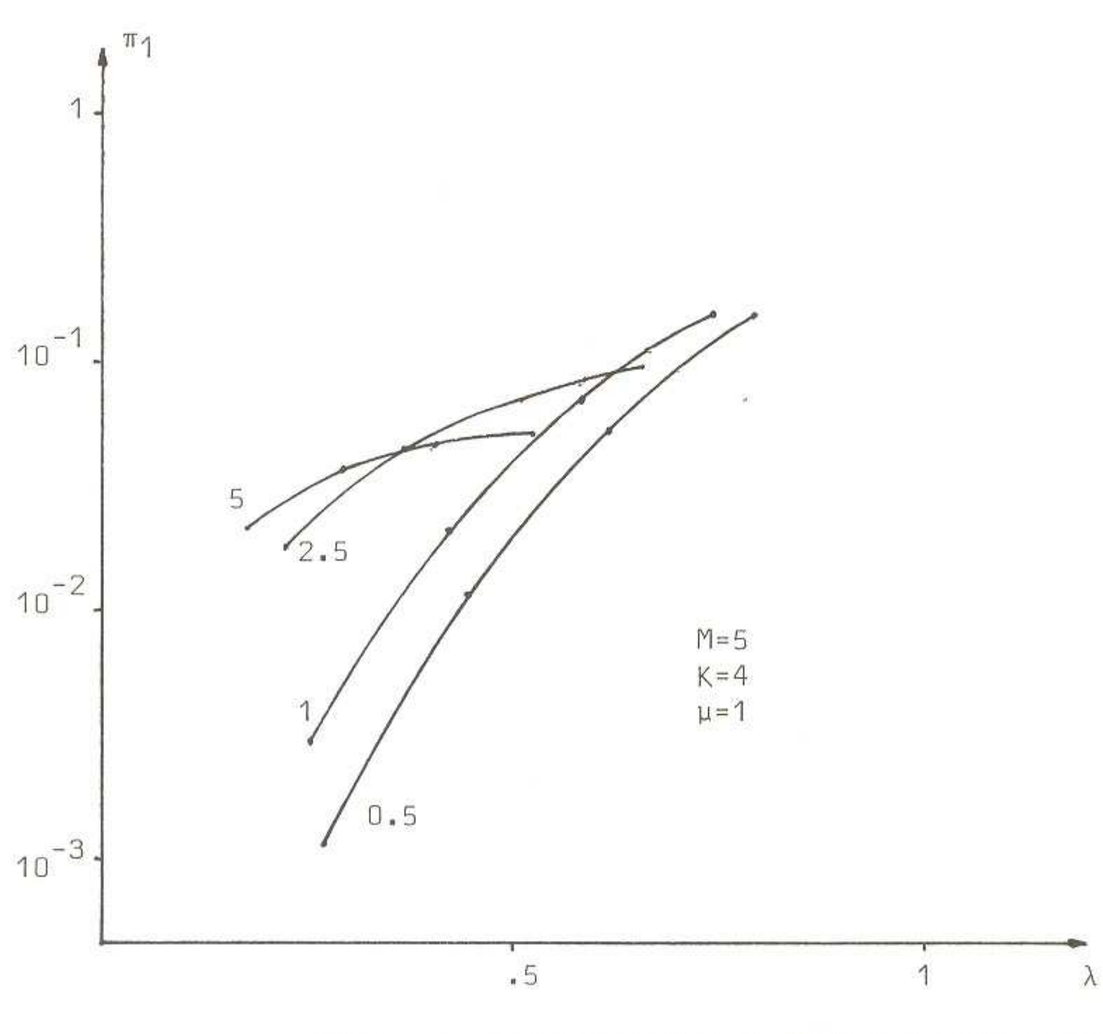


Figure 6. Probability that Unit I becomes blocked

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

ENDEMICITE DANS LE MODELE D'EPIDEMIE LOGISTIQUE A TEMPS DISCRET

Claude LEFEVRE

Université Libre de Bruxelles
Institut de Statistique, C.P. 210
Boulevard du Triomphe
1050 Bruxelles
Belgique

ABSTRACT

This paper is concerned with the problem of endemicity in the deterministic version of the discrete time logistic epidemic model. Conditions for endemicity are first derived, explicit bounds for the endemic level are then constructed, and the effects of small perturbations on some control parameters are finally analyzed.

1. Introduction

Le modèle à temps discret d'épidémie logistique concerne une population fermée de N individus, homogène et constamment brassée, et subdivisée en deux classes : les susceptibles, individus susceptibles de contracter la maladie, et les infectés, individus porteurs du germe infectieux et qui le transmettent. L'état de la population est observé aux instants $t = 0,1,2,\ldots$. Notons $i(t)$ et $s(t)$ respectivement les nombres d'infectés et de susceptibles à l'instant t. Comme la population est fermée, $s(t) + i(t) = N$ pour tout t. Supposons qu'il y ait des infectés présents dans la population à l'instant initial, c'est-à-dire, $0 < i(0) \leq N$. L'évolution de l'épidémie pendant l'intervalle de temps $(t, t+1]$ est régie par les deux processus suivants.

a) Chacun des $i(t)$ infectés a la probabilité g de guérir et redevenir alors susceptible à l'instant t+1. Nous supposons $0 < g \leq 1$.

b) Chacun des $s(t)$ susceptibles devient infecté à l'instant t+1 s'il a au moins un contact infectieux pendant l'intervalle $(t, t+1]$. Le calcul de la probabilité d'infection d'un susceptible est présenté à la section 2.

Dans ce travail, nous nous intéressons uniquement à la version déterministe du modèle. Nous construisons cette version à la section 3. Dans la section 4, nous déterminons les conditions conduisant à une situation endémique, et nous dérivons ensuite des bornes explicites pour le niveau endémique. Enfin, nous étudions dans la section 5 les effets de variations locales des paramètres de contrôle sur le niveau endémique.

2. Le processus de propagation de l'infection

Pour décrire la manière dont l'infection se propage, nous suivons une suggestion de Dietz et Schenzle [3] en tenant compte explicitement de la distribution du nombre de contacts entre individus. Soit R la variable aléatoire représentant le nombre de rencontres que peut faire un susceptible donné pen-

dant une unité de temps. La variable R est supposée indépendante du nombre d'infectés présents. Nous notons $G(z)$, $0 \leq z \leq 1$, la fonction génératrice de R, et nous faisons l'hypothèse, peu restrictive en pratique, $0 < E(R^2) < \infty$.

Plaçons-nous à l'instant t; il y a alors i(t) infectés dans la population. Pour simplifier l'écriture, nous omettons l'argument t dans cette section. Soit $y = i/N$ la proportion d'infectés présents. Nous supposons que lors d'un contact avec un susceptible, un infecté transmet le germe infectieux avec la probabilité φ, $0 < \varphi \leq 1$. Par conséquent, un susceptible qui rencontre un individu de la population deviendra infecté suite à ce contact avec la probabilité φy.

Notons C la variable aléatoire représentant le nombre total de contacts infectieux faits par un susceptible donné pendant une unité de temps. Pour l'intervalle $(t, t+1]$, on a

$$C = \sum_{j=1}^{R} x_j$$

où les x_j sont des variables aléatoires indépendantes et distribuées suivant la loi de Bernoulli de paramètre φy. Dès lors, la fonction génératrice conditionnelle de $C|y$ est donnée par

$$E(z^C|y) = G(1 - \varphi y + \varphi y z), \quad 0 \leq z \leq 1 . \tag{1}$$

Remarque. Supposons que R ait une distribution $\mathcal{P}(\theta)$ dépendant d'un paramètre $\theta \in \Theta$, et soit $G(z;\theta)$ sa fonction génératrice. Suivant la définition de Berg [1], la famille de distributions $\{\mathcal{P}(\theta), \theta \in \Theta\}$ est dite fermée binomialement si pour tous $\theta \in \Theta$ et $\gamma \in [0,1]$, il existe un $\tilde{\theta}(\theta,\gamma) \in \Theta$ tel que

$$G(1 - \gamma + \gamma z; \theta) = G[z; \tilde{\theta}(\theta,\gamma)], \quad 0 \leq z \leq 1 .$$

Si c'est le cas, nous déduisons de (1) que

$$E(z^C|y) = G[z; \tilde{\theta}(\theta, \varphi y)], \quad 0 \leq z \leq 1 ,$$

c'est-à-dire que $C|y$ a comme distribution $\mathcal{P}[\tilde{\theta}(\theta, \varphi y)]$. Ainsi par exemple, les familles de distributions suivantes sont binomialement fermées : {Poisson (λ), $\lambda > 0$}, {Binomiale (M,p), $0 < p < 1$}, {Binomiale négative (M,p), $0 < p < 1$},

et on trouve alors que

$$\begin{cases} \text{si } R \sim \text{Poisson } (\lambda), \text{ alors } C|y \sim \text{Poisson } (\lambda\varphi y) , \\ \text{si } R \sim \text{Binomiale } (M,p), \text{ alors } C|y \sim \text{Binomiale } (M, p\varphi y) , \\ \text{si } R \sim \text{Binomiale négative } (M,p), \text{ alors } C|y \sim \text{Binomiale négative} \\ (M, p/[\, p + (1-p)\varphi y]\,) . \end{cases}$$

Finalement, faisons l'hypothèse habituelle qu'un susceptible devient infecté s'il a au moins un contact infectieux. De (1), nous obtenons que la probabilité qu'il contracte la maladie pendant (t, t+1] est égale à

$$P_1(y) = P(C \geq 1|y)$$
$$= 1 - G(1 - \varphi y) . \qquad (2)$$

3. Le modèle dépidémie logistique

Comme expliqué à la section 1, le modèle d'épidémie logistique tient compte de deux processus, la guérison d'infectés et l'infection de susceptibles. Dans la version déterministe du modèle, le nombre i(t+1) d'infectés en t+1 est défini par la relation de récurrence

$$i(t+1) = (1-g)\, i(t) + P_1[y(t)][N - i(t)]$$
$$= (1-g)\, i(t) + \{1 - G[\,1 - \varphi y(t)]\}[N - i(t)] \quad . \qquad (3)$$

Divisons (3) par N. Nous obtenons que $y(t) = i(t)/N$ est solution de l'équation aux différences du premier ordre

$$y(t+1) = f[\,y(t)] \ , \qquad (4a)$$

où

$$f[\,y(t)] = (1-g)\, y(t) + \{1 - G[\,1 - \varphi y(t)]\}[\,1 - y(t)] \ . \qquad (4b)$$

4. Le phénomène d'endémicité

Nous commençons par déterminer les conditions conduisant à une situation endémique au sein de la population. En adaptant la méthode de Cooke, Calef et Level [2], on peut démontrer le théorème 1 ci-dessous. Notons $m_1 = E(R)$.

Théorème 1. Si $m_1 \varphi \leq g$, alors f(y) a un seul point fixe, 0. Si $m_1 \varphi > g$, alors f(y) a deux points fixes, 0 et y*, ce dernier étant la racine positive de l'équation

$$y = (1/g)\{1 - G[1 - \varphi y]\}(1-y) \; ; \qquad (5)$$

l'état 0 est instable tandis que l'état y* est globalement stable.

Dans le contexte épidémiologique, ce théorème montre que si $m_1 \varphi \leq g$, l'épidémie finira par s'éteindre tandis que si $m_1 \varphi > g$, l'épidémie deviendra endémique : c'est le phénomène de seuil. La condition $m_1 \varphi \lessgtr g$ est intuitive et consiste à comparer les nombres moyens de contacts infectieux et de guérisons pendant une unité de temps où la proportion d'infectés présents est infiniment petite.

Envisageons la situation où l'épidémie devient endémique, c'est-à-dire, supposons $m_1 \varphi > g$. Le niveau endémique y* est alors la racine positive de (5). Le plus souvent, y* ne peut être calculé explicitement et il est donc intéressant de construire des bornes simples pour y*. Le théorème 2 ci-dessous fournit des bornes explicites pour y* qui reposent sur la seule connaissance de quelques paramètres importants de la distribution de R.

Théorème 2.

. Notons $m_2 = E[R(R-1)]$ et définissons

$$\begin{cases} k = \text{partie entière de } (m_1 + m_2)/m_1 \; , & (6a) \\ c = 2\, m_1/(k+1) - m_2/k(k+1) \; . \end{cases}$$

Alors une borne inférieure pour y* est $y_{inf.}$ donnée par

$$y_{inf.} = (2/\varphi)\{1 - [(c + g + m_1(1-\varphi))/(c + m_1)]^{1/(k+1)}\} \; . \qquad (6b)$$

. Notons $p_o = P(R = 0)$ et définissons

$$\begin{cases} \ell = \text{partie entière de } m_1/(1 - p_o) \; , & (7a) \\ \eta = m_1 - \ell(1 - p_o) \; . \end{cases}$$

Alors une borne supérieure pour y* est $y_{sup.}$ donnée par

$$y_{sup.} = 1 - g/[(1-p_o+g) - (1-p_o-\eta)(1-\varphi)^{\ell} - \eta(1-\varphi)^{\ell+1}] \; . \qquad (7b)$$

La démonstration de ce théorème est assez longue et n'est pas reprise ici; une version détaillée peut être obtenue auprès de l'auteur. Il convient de souligner que $y_{inf.}$ dépend uniquement des moments m_1 et m_2, et que $y_{sup.}$ dépend uniquement de la moyenne m_1 et de la probabilité p_0.

5. *Effets de variations locales de paramètres de contrôle*

Plaçons-nous dans le cas où l'épidémie a atteint un niveau d'endémicité $y^* > 0$. Les deux politiques de santé publique usuelles consistent d'une part à améliorer la qualité des soins apportés aux malades, et d'autre part à renforcer les mesures préventives pour les personnes en bonne santé. Dans le modèle d'épidémie logistique, elles correspondent, respectivement, à augmenter g, la probabilité qu'un infecté guérisse pendant une unité de temps, et à augmenter $1-\varphi$, la probabilité que lors d'un contact avec un infecté, un susceptible ne reçoive pas le germe infectieux et reste en bonne santé.

Il est intuitif, et on le démontre facilement, que si g ou $1-\varphi$ augmentent, alors y^* diminue. En pratique, les modifications qui peuvent être apportées à ces paramètres sont généralement, à court et moyen terme, de très faible amplitude. De plus, il arrive souvent que pour des raisons médicales ou matérielles, une seule de ces deux politiques puisse être mise en oeuvre. Il convient alors de déterminer la "meilleure stratégie locale", c'est-à-dire, celle qui, par une légère variation du paramètre, diminue le plus sensiblement le niveau d'endémicité y^*.

Une augmentation locale de g sera préférable à une augmentation locale de $1-\varphi$ lorsque

$$|\partial y^*/\partial g| \geq |\partial y^*/\partial(1-\varphi)| \quad . \tag{8}$$

De (5), nous obtenons que (8) s'écrit encore

$$1 \geq (1-y^*)\, G'(1 - \varphi y^*) \quad . \tag{9}$$

Il est clair que si $m_1 \leq 1$, alors l'inégalité (9) est toujours satisfaite. En d'autres termes, si le nombre moyen de rencontres par unité de temps ne dépasse

pas 1, la meilleure des deux politiques est d'augmenter localement g; ce résultat n'est pas surprenant. Si, au contraire, $m_1 > 1$, alors (9) n'est plus nécessairement satisfait, et il peut être préférable d'augmenter localement $1-\varphi$.

A titre d'illustration, nous continuons la discussion de ce problème dans les cas particuliers où $R \sim$ Géométrique (p) et $R \sim$ Poisson (λ), avec $m_1\, \varphi > g$. On peut montrer qu'augmenter localement $g[1-\varphi]$ est préférable si $1-\varphi$ est inférieur [supérieur] à une valeur critique $1-\varphi(g)$, où $\varphi(g)$ est définie pour la distribution géométrique par

$$\varphi(g) = [1/2(1-p)]\{-p + [p^2 + 4\, p(1-p)g(g+1)]^{1/2}\}\ , \qquad (10)$$

et pour la distribution de Poisson par

$$\varphi(g) = [(g+1)/(\lambda-1)]\ \ln[(g\lambda + 1)/(g+1)]\ . \qquad (11)$$

Dans (10) et (11), $\varphi(g)$ est une fonction croissante, concave et telle que $\varphi(0) = 0$, $\varphi(1) < 1$. Ce résultat nous apprend donc que la politique optimale consiste à augmenter localement $g[1-\varphi]$ si g et $1-\varphi$ sont suffisamment petits [grands]. En gros, il convient de traiter les infectés ou les susceptibles selon que la maladie est dure ou bénigne.

BIBLIOGRAPHIE

[1] Berg, M., Stochastic Models for Spread of Motivating Information, Naval Res. Log. Quart., 28, 1981, nr. 1, pp.133-145.

[2] Cooke, K.L., Calef, D.F. and Level, E.V., Stability or Chaos in Discrete Epidemic Models, in Nonlinear Systems and Applications - An International Conference, 1977, Academic, New York, pp.75-93.

[3] Dietz, K. and Schenzle, D., Mathematical Models for Infectious Disease Statistics, in A.C. Atkinson and S.E. Fienberg (eds.), A Celebration of Statistics, 1985, Springer, New York, pp.167-204.

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

ANALYSIS OF SINGLE SERVER QUEUEING SYSTEMS WITH VACATION PERIODS

Jacqueline LORIS-TEGHEM

Université de l'Etat à Mons (Belgium)

ABSTRACT

For an M/G/1-type queueing system with a different service time distribution for the first customer served in a busy period, we consider two types of vacation policies for the removable server and investigate the transient and steady-state behaviour of the waiting time process for the FIFO discipline. We then extend the steady-state analysis to the case where the server applies a combination of a vacation policy and the (0,k)-policy.

1. *Introduction*

In [1], Levy & Yechiali studied the steady-state of two M/G/1 models in which the removable server leaves the system for a "vacation period" whenever a service terminates with no customers left in the queue.

We show that "Model 1" and "Model 2" in [1] - extended by considering a different service time distribution for the first customer served in a busy period - are both examples of the generalized queueing system considered in [2] and [3], for which the transient behaviour of the waiting time process was studied via an algebraic approach based on the concept of Wendel projection in the case of arbitrary interarrival and service time distributions [2] and by using integral representations of the involved operators in the case of interarrival times or service times having a rational characteristic function [3]. From the results in [3], we derive the transient behaviour of the waiting time process for our extended Models 1 and 2 and, as a limit result, we get the stationary distribution of this process.

We then generalize the models by considering a combination of the vacation policy and the (0,k)-policy and extend the steady-state analysis of [1] for what the queue length concerns.

2. *Description of the models*

Customers $\mathcal{C}_n$, $n \geqslant 0$, arrive according to a Poisson process of parameter λ and are served in the order of their arrival.

We introduce the following notations relative to customer $\mathcal{C}_n$ $(n \geqslant 0)$:

T_n : the arrival instant ;

T'_n : the service initiation instant ;

T''_n : the departure instant

and define the random variables :

$a_n = T_{n+1} - T_n$ (interarrival time) ;

$d_n = T'_n - \max(T''_{n-1}, T_n)$ (delay imposed on $\mathcal{C}_n$) ;

$s_n = T''_n - T'_n$ (service time of $\mathcal{C}_n$) ;

$w_n = T'_n - T_n$ (waiting time of $\mathcal{C}_n$) ;

$v_n = T''_n - T_n$ (sojourn time of $\mathcal{C}_n$),

for which the following relations hold :

$$v_n = [v_{n-1} - a_{n-1}]^+ + d_n + s_n \qquad (n \geqslant 1)$$

$$w_n = [v_{n-1} - a_{n-1}]^+ + d_n$$

For both models, we have (for $n \geqslant 1$) :

$$s_n = {}_0s_n \quad \text{if } T_n \leqslant T''_{n-1} ;$$

$$= {}_1s_n \quad \text{if } T_n > T''_{n-1}$$

where $\{({}_0s_n, {}_1s_n)\}_{n\geqslant 1}$ is a sequence of i.i.d. random vectors, independent of $\{a_n\}_{n\geqslant 0}$.

For Model 1 :

$$d_n = 0 \quad \text{if } T_n \leqslant T''_{n-1} \text{ or } T_n > T''_{n-1} + u_n ;$$

$$= T''_{n-1} + u_n - T_n \quad \text{if } T''_{n-1} < T_n \leqslant T''_{n-1} + u_n,$$

where $\{u_n\}_{n\geqslant 1}$ is a sequence of i.i.d. random variables, independent of $\{({}_0s_n, {}_1s_n, a_{n-1})\}_{n\geqslant 1}$.

(u_n is the duration of the single vacation during which the server leaves the system if no customers are left in the queue at T''_{n-1})

For Model 2 :

$$d_n = 0 \qquad \text{if } T_n \leqslant T''_{n-1} \ ;$$

$$T''_{n-1} + \sum_{\nu=1}^{i} u_{n,\nu} - T_n \qquad \text{if } T''_{n-1} + \sum_{\nu=1}^{i-1} u_{n,\nu} < T_n \leqslant T''_{n-1} + \sum_{\nu=1}^{i} u_{n,\nu}$$

$(i \geqslant 1)$

where the $u_{n,\nu}$ $(n \geqslant 1,\ \nu \geqslant 1)$ are i.i.d. random variables independent of $\{({}_0s_n,\ {}_1s_n,\ a_{n-1})\}_{n\geqslant 1}$.

$(u_{n,1},\ u_{n,2},\ \ldots$ are the durations of the successive vacations during which the server leaves the system if no customers are left in the queue at T''_{n-1}, until, coming back from such a vacation, he finds a non empty queue)

3. *Transient and steady-state behaviour of the sojourn time and the waiting time processes*

Using the fact that the arrival process is a Poisson process, it can be derived from the above description that :

$$d_n = 0 \quad \text{if} \quad T_n \leqslant T''_{n-1} \ ;$$

$${}_1d_n \quad \text{if} \quad T_n > T''_{n-1},$$

where $\{{}_1d_n\}_{n\geqslant 1}$ is a sequence of i.i.d. random variables independent of $\{({}_0s_n,\ {}_1s_n,\ a_{n-1})\}_{n\geqslant 1}$, the common distribution of which is given by :

for Model 1 :

$$E\,(e^{-\theta\ {}_1d_n}) = \frac{1}{\theta - \lambda}\,[\theta\ u\,(\lambda) - \lambda\ u\,(\theta)] \qquad (1)$$

for Model 2 :

$$E\,(e^{-\theta\ {}_1d_n}) = \frac{1}{1 - u\,(\lambda)}\,\frac{\lambda}{\theta - \lambda}\,[u\,(\lambda) - u\,(\theta)], \qquad (2)$$

where $u\,(\theta)$ denotes either $E\,(e^{-\theta\ u_n})$ or $E\,(e^{-\theta\ u_{n,\nu}})$.

Thus either model is an example of the generalized queueing system considered in [2] and [3].

Supposing that v_o is independent of the vectors $({}_1d_n, {}_os_n, {}_1s_n, a_{n-1})$, $n \geqslant 1$, and using the following notations :

$$h_o(\theta) = E(e^{-\theta v_o}) \; ; \; g_o(\theta) = E(e^{-\theta w_o})$$

$$h_{(z)}(\theta) = \sum_{n \geqslant 0} z^n E(e^{-\theta v_n}) \; ; \; g_{(z)}(\theta) = \sum_{n \geqslant 0} z^n E(e^{-\theta w_n}) \qquad (|z| < 1)$$

$${}_1d(\theta) = E(e^{-\theta\, {}_1d_n}) \; ; \; {}_is(\theta) = E(e^{-\theta\, {}_is_n}) \qquad (i = 1, 2),$$

we deduce from the results in [3] that :

$$h_{(z)}(\theta) = \frac{1}{\theta - \lambda + z\lambda\, {}_os(\theta)} \{(\theta - \lambda) h_o(\theta) + z\, \xi_{(z)} [\lambda\, {}_os(\theta) + (\theta - \lambda)\, {}_1d(\theta)\, {}_1s(\theta)]\}$$

$$g_{(z)}(\theta) = \frac{1}{\theta - \lambda + z\lambda\, {}_os(\theta)} \{(\theta - \lambda) g_o(\theta) + \lambda z [{}_os(\theta)\, g_o(\theta) - h_o(\theta)] + z\xi_{(z)} [\lambda + (\theta - \lambda)\, {}_1d(\theta) + z\lambda\, {}_1d(\theta) ({}_os(\theta) - {}_1s(\theta)]\}$$

with $\xi_{(z)} = \dfrac{h_o(\varepsilon(z))}{1 - z\, {}_1d(\varepsilon(z))\, {}_1s(\varepsilon(z))}$

where, for $|z| < 1$, $\varepsilon(z)$ denotes the unique zero of the function $\theta - \lambda + z\lambda\, {}_os(\theta)$ in the half-plan Re $\theta > 0$.
The function ${}_1d(\theta)$ is given by (1) and (2) for Models 1 and 2 respectively.

Putting $\bar{u} = E(u_n) = E(u_{n,\nu})$

$${}_1\bar{d} = E({}_1d_n) \; ; \; {}_i\bar{s} = E({}_is_n) \quad (i = 1, 2)$$

and supposing that these expectations are finite and that $\lambda\ {}_o\bar{s} < 1$, one gets, for the stationary distributions

$$h(\theta) = \lim_{z \to 1} (1 - z)\ h_{(z)}(\theta)\ ;\ g(\theta) = \lim_{z \to 1} (1 - z)\ g_{(z)}(\theta)\ :$$

$$h(\theta) = \xi\ \frac{\lambda\ {}_o s(\theta) + (\theta - \lambda)\ {}_1 d(\theta)\ {}_1 s(\theta)}{\theta - \lambda + \lambda\ {}_o s(\theta)} \qquad (3)$$

$$g(\theta) = \xi\ \frac{\lambda + (\theta - \lambda)\ {}_1 d(\theta) + \lambda\ {}_1 d(\theta)\ [{}_o s(\theta) - {}_1 s(\theta)]}{\theta - \lambda + \lambda\ {}_o s(\theta)}$$

with $\xi = \dfrac{1 - \lambda\ {}_o\bar{s}}{1 - \lambda\ ({}_o\bar{s} - {}_1\bar{s}) + \lambda\ {}_1\bar{d}}$

For Model 1, ${}_1 d(\theta)$ is given by (1) and ${}_1\bar{d} = \dfrac{\lambda\ \bar{u} + u(\lambda) - 1}{\lambda}$

For Model 2, ${}_1 d(\theta)$ is given by (2) and ${}_1\bar{d} = \dfrac{\lambda\ \bar{u} + u(\lambda) - 1}{\lambda\ (1 - u(\lambda))}$

Particularizing relation (3) to the case where ${}_o s(\theta) \equiv {}_1 s(\theta)$, one gets the expressions given in [1] for the steady-state distribution of the sojourn time in Models 1 and 2.

4. Combination of the vacation policies with the (0,k)-policy

In this section, we generalize the models described in section 2 by combining the vacation policy with the (o,k)-policy ($k \geqslant 1$) i.e. :

for Model 1 : when a service terminates with no customers left in the system, the server leaves for a single vacation. When coming back, he immediately initiates a busy period if at least k customers are queueing ; otherwise, he waits until k customers are present to start serving again ;

for Model 2 : when a service terminates with no customers left in the system, the server leaves for successive vacations, until, coming back from such a vacation, he finds at least k customers queueing.

The steady-state analysis performed in [1] for k = 1 can be readily extended. We consider the instants $\tau_1, \tau_2, \ldots, \tau_n, \ldots$ at which either a service or a vacation period terminates (where by vacation period, we mean a single vacation for Model 1 and a sequence of successive vacations for Model 2) and we define $x_n = (i_n, l_n)$, where l_n denotes the queue length at $\tau_n + 0$ and i_n is 0 (respectively 1) if a vacation period (respectively a service) terminates at τ_n.

Putting

$$p_{j|1}^{(k)} = \lim_{n\to\infty} \Pr [l_n = j | i_n = 1]$$

and supposing that $\lambda \, {}_o\bar{s} < 1$, we get the following expression for the generating function $P_{|1}^{(k)}(z) = \sum_{j\geqslant 0} z^j \, p_{j|1}^{(k)}$ $(|z| \leqslant 1)$:

- for Model 1

$$P_{|1}^{(k)}(z) = \xi_1^{(k)} \frac{{}_1s\,(\lambda(1-z))\,[u(\lambda(1-z)) + \sum_{r=0}^{k-1} b_r\,(z^k - z^r)] - {}_os\,(\lambda(1-z))}{z - {}_os\,(\lambda(1-z))}$$

$$\text{with } \xi_1^{(k)} = \frac{1 - \lambda \, {}_o\bar{s}}{\lambda\,\bar{u} + \sum_{r=0}^{k-1} b_r\,(k-r) - \lambda\,({}_o\bar{s} - {}_1\bar{s})}$$

$$\text{where } b_r = \int_0^\infty e^{-\lambda t} \frac{(\lambda t)^r}{r!} \, d \Pr [u_n \leqslant t] \qquad (r \geqslant 0)$$

- for Model 2

$$P_{|1}^{(k)}(z) = \xi_2^{(k)} \frac{{}_1s\,(\lambda(1-z))\,\tilde{B}^{(k)}(z) - {}_os\,(\lambda(1-z))}{z - {}_os\,(\lambda(1-z))} \qquad (4)$$

$$\text{with } \xi_2^{(k)} = \frac{1 - \lambda \, {}_o\bar{s}}{\overline{\tilde{n}^{(k)}} - \lambda\,({}_o\bar{s} - {}_1\bar{s})}$$

$$\text{where } \tilde{B}^{(k)}(z) = \sum_{r\geqslant k} z^r \Pr [\tilde{n}^{(k)} = r]$$

$$\overline{\tilde{n}^{(k)}} = E(\tilde{n}^{(k)}),$$

$\tilde{n}^{(k)}$ denoting the number of customers arriving during a vacation period.

The $\tilde{B}^{(k)}(z)$ and $\overline{\tilde{n}^{(k)}}$, $k \geqslant 1$, can be recursively derived using the following relations :

$$\tilde{B}^{(k)}(z)\,(1 - b_o) = \sum_{r=1}^{k-1} b_r z^r \left[\tilde{B}^{(k-r)}(z) - 1\right] + u\left(\lambda(1-z)\right) - b_o \qquad (5)$$

$$\overline{\tilde{n}^{(k)}}\,(1 - b_o) = \sum_{r=1}^{k-1} b_r \overline{\tilde{n}^{(k-r)}} + \lambda \bar{u} \qquad (6)$$

Remark : a further extension

Model 2 can be further extended by allowing the distribution of the duration $u_{n,\nu}$ of the ν^e vacation in a vacation period to depend on the number of customers present in the system when the vacation begins : given that this number is equal to r ($r = 0, \ldots, k-1$), the conditional distribution of $u_{n,\nu}$ has the Laplace-Stieltjes transform $u_r(\theta)$ (we put $u_0(\theta) \equiv u(\theta)$).
[This generalization was suggested to me by Jacques Teghem Jr.]

Relation (4) still applies, as well as relations (5) and (6), the random variable $\tilde{n}^{(k-r)}$ corresponding now to the model combining the (0, k-r)-policy and the vacation policy in which the conditional distribution of a vacation, given that r' customers are present when it begins, has Laplace-Stieltjes transform $u_{r'+r}(\theta)$ ($r' = 0, \ldots, k-r-1$).

REFERENCES

[1] LEVY, Y. and YECHIALI, U. : "Utilization of idle time in an M/G/1 queueing system", Man. Science, 22, pp. 202-211 (1975).

[2] LORIS-TEGHEM, J. : "On the waiting time distribution in a generalized GI/G/1 queueing system", J. Appl. Prob., 8, pp. 241-251 (1971).

[3] LORIS-TEGHEM, J. : "On the waiting time distribution in some generalized queueing systems", Bull. Soc. Math. Belg., XXV, pp. 11-24 (1973).

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

PROCESSUS DES PERIODES D'OCCUPATION D'UN MODELE D'ATTENTE DU TYPE $M_n/M_n/1$

MANYA NDJADI

Université de Kinshasa,
B.P. 190, Kinshasa XI, Zaïre

RESUME

Nous considérons un modèle d'attente du type processus de vie et de mort homogène: $\lambda_n = (n + 1)\lambda$ et $\mu_n = n\,\mu$; $\lambda > 0$, $\mu > 0$.
Il s'agit d'étudier le processus aléatoire $\{T_k, k \geqslant 1\}$. Par définition T_k est la longueur d'un intervalle de temps qui commence à tout instant où le système contient k clients et finit à l'instant où le système devient vide pour la première fois.

ABSTRACT

We consider a queueing system generated by an homogenous birth and death process of following type:
$\lambda_n = (n + 1)\lambda$ and $\mu_n = n\,\mu$; $\lambda > 0$, $\mu > 0$.
We study the randon process $\{T_k, k \geqslant 1\}$ where T_k denotes the length of a time interval starting at each instant when the system contains k customers and ending at the instant when the system becomes empty for the first time.

1. *Introduction*

Les modèles d'attente du type processus de vie et de mort sont représentés, en général, par la notation $M_n/M_n/1$. Cette notation rappelle que les paramètres taux d'arrivée λ_n et taux de service μ_n varient avec l'état du système. C'est une généralisation dynamique du modèle statique M/M/1. Parmi les modèles d'attente du type $M_n/M_n/1$, nous pouvons citer :

Modèle A : ($\lambda_n = \lambda$, $\mu_n = n\mu$) ;

Modèle B : ($\lambda_n = (n + 1)\lambda$, $\mu_n = n\mu$) ;

Modèle C : ($\lambda_n = \frac{\lambda}{n+1}$, $\mu_n = \mu$).

Ce classement alphabétique suit l'ordre croissant des difficultés dans le calcul des probabilités d'état en régime transitoire (R.T.), [4].

Le problème des périodes d'occupation en R.T., qui nous intéresse ici, concerne le Modèle B. Notons que ce Modèle B rappelle un centre de service où un afflux de demandes provoque une réaction compensatoire du serveur. C'est surtout dans ce modèle que s'exprime cette généralisation dynamique du modèle statique M/M/1, [1 ; 4].

Le processus des périodes d'occupation en R.T., concerne en général, deux variables aléatoires (V.A.) :

- la longueur d'un intervalle de temps T_k, ($k \geq 1$), qui commence à tout instant où le système - nombre de clients - se trouve dans l'état k et finit à l'instant où le système devient vide pour la première fois ;
- le nombre $N(T_k)$ de clients servis durant T_k.

Nous considérons en particulier la première de ces deux V.A. C'est-à-dire, en notant par $\gamma_{kr}(t)dt = IP\left[t < T_k < t + dt, N(T_k) = r\right]$, ($r \geq k$) (1) et par $\Gamma_k(x, t)$ la fonction génératrice des $\gamma_{kr}(t)$, nous nous intéressons à l'expression de $\Gamma_k(1, t)$ pour le Modèle B. $\Gamma_k(1, t)$ est la densité de probabilité de la période d'occupation T_k, ($k \geq 1$). Nous supposons qu'à l'instant t = 0 le système contient k clients et que les V.A. T_k sont stochastiquement indépendantes et équidistribuées.

HADIDI, [3], s'est occupé de ce problème mais dans le cas des Modèles A et C uniquement. Il a calculé, pour chacun de ces deux modèles, la transformée de Laplace de la densité de probabilité de la V.A. T_k. D'où l'on peut, en principe, obtenir tous les moments de cette V.A. T_k. Nous montrons que cette méthode de HADIDI peut s'étendre au Modèle B.

2. *Equation fonctionnelle régissant le processus des périodes d'occupation dans le modèle B*

La fonction $\gamma_{kr}(t)$, $(r \geqslant k \geqslant 1)$, définie par la relation (1), est la densité de probabilité d'une période d'occupation T_k au cours de laquelle r clients sont servis.

Soient $\gamma^*_{kr}(z)$ la transformée de Laplace de $\gamma_{kr}(t)$ et $\Gamma^*_k(x, z)$ celle de la fonction génératrice

$$\Gamma_k(x, t) = \sum_{\ell=0}^{\infty} x^{k+\ell} \gamma_{k,\, k+\ell}(t) , \quad |x| \leqslant 1. \tag{2}$$

En procédant comme HADIDI [3] pour les Modèles A et C, nous obtenons l'équation fonctionnelle suivante pour le Modèle B :

$$\begin{aligned} k\lambda\Gamma^*_k(x,z) = &\big((z + k\lambda + (k-1)\mu\big)\Gamma^*_{k-1}(x,z) \\ &- (k-1)\mu x\, \Gamma^*_{k-2}(x,z) \end{aligned} \tag{3}$$

$(k \geqslant 2)$.

L'équation (3) régit tout le processus des périodes d'occupation du Modèle B. Sa résolution en x=1 fournit l'expression de $\Gamma^*_k(1,z)$; d'où l'on pourra calculer tous les moments de la V.A. T_k. En effet, $\Gamma^*_k(1,z)$ étant la transformée de Laplace de la densité de probabilité $\Gamma_k(1,t)$ de T_k, l'on se souvient alors que

$$m_n^{(k)} = \mathbb{E}\,(T_k^n) = (-1)^n \left.\frac{d^n\Gamma^*_k(1,z)}{dz^n}\right|_{z\,=\,0} \tag{4}$$

3. *Calcul de l'expression de* $\Gamma_k^*(1,z)$.

En x=1, l'équation (3) donne

$$k\lambda\ \Gamma_k^*(1,z) = \left[z + k\lambda + (k-1)\mu\right]\Gamma_{k-1}^*(1,z) - (k-1)\mu\Gamma_{k-2}^*(1,z). \tag{5}$$

En réalité, (5) est un système infini d'équations algébriques. D'où une méthode adéquate pour le résoudre est de recourir à la fonction génératrice

$$L(1,z,y) = \sum_{k=1}^{\infty} y^k\ \Gamma_k^*(1,z),\ |y| < 1 \quad \text{(au sens strict)}. \tag{6}$$

Dès lors, (5) équivaut à l'équation différentielle linéaire en la variable y (on considère z comme paramètre) :

$$\frac{\partial L}{\partial y} + \frac{\mu y - (z+\lambda)}{(1-y)(\lambda-\mu y)} L = \frac{\lambda\Gamma_1^*(1,z)-\mu y}{(1-y)(\lambda-\mu y)}\ ,\quad |y|<1\ ;$$

$$\text{CI : } L(1,z,0) = 0. \tag{7}$$

La solution de (7), compte tenu de CI, est alors :

$$L(1,z,y) = \left[(\lambda-\mu y)/\lambda\right]^{\alpha}\left[\Gamma_1^*(1,z)\sum_{n=o}^{\infty}\frac{Q_n(\alpha)y^n}{n!(\lambda/\mu)^n}\right.$$

$$\times \sum_{j=o}^{\infty}\frac{n!\ H_j(\alpha)y^{j+1}}{(n+j+1)!\ (1-y)^{j+1}}$$

$$\left.-\sum_{n=o}^{\infty}\frac{Q_n(\alpha)y^{n+1}}{(n!\ (\lambda/\mu)^{n+1}}\sum_{j=o}^{\infty}\frac{(n+1)!\ H_j(\alpha)y^{j+1}}{(n+j+2)!\ (1-y)^{j+1}}\right] \tag{8}$$

Les fonctions auxiliaires introduites dans la formule (8) sont ainsi définies :

$$H_j(\alpha) = \begin{cases} o & \text{si } j \notin \mathbb{N}, \\ 1 & \text{si } j = o, \\ \alpha(\alpha-1)\ldots(\alpha-j+1) & \text{si } j \geqslant 1\ ; \end{cases}$$

$$Q_j(\alpha) = \begin{cases} o & \text{si } j \notin \mathbb{N}, \\ 1 & \text{si } j = o \\ (\alpha+1)\ldots(\alpha+j) & \text{si } j \geqslant 1\ ; \end{cases}$$

où $\alpha = z/(\lambda-\mu)$.

L'expression de L(1, z, y) ci-dessus se met finalement sous la forme de séries entières en y ; séries qui convergent absolument et uniformément pourvu que $|y| < \frac{\lambda}{\mu} < 1$.

$$L(1, z, y) = \sum_{\nu=o}^{\infty} \frac{(-1)^{\nu} H_{\nu}(\alpha) y^{\nu}}{\nu! (\lambda/\mu)^{\nu}} \sum_{n=o}^{\infty} \frac{Q_{n}(\alpha)}{n! (\lambda/\mu)^{n+1}}$$

$$x \{\frac{\lambda}{\mu} \Gamma_1^*(1, z) \sum_{j=o}^{\infty} \frac{n! H_j(\alpha) y^{n+j+1}}{(n+j+1)!} \sum_{r=o}^{\infty} \binom{j+r}{j} y^{r}$$

$$- \sum_{j=o}^{\infty} \frac{(n+1)! H_j(\alpha) y^{n+j+2}}{(n+j+2)!} \sum_{r=o}^{\infty} \binom{j+r}{j} y^{r} \} \qquad (9)$$

Par la définition (6), $\Gamma_k^*(1, z)$ est le coefficient de y^k dans (9).
Et en vertu de la convergence absolue, le calcul du terme en y^k se fait comme dans une somme d'un nombre fini de termes. Il en résulte que :

$$\Gamma_k^*(1, z) = \sum_{\nu=o}^{k-1} \frac{(-1)^{\nu} H_{\nu}(\alpha)}{\nu!(\lambda/\mu)^{\nu}} \Gamma_1^*(1, z) \sum_{n=o}^{k-\nu-1} \frac{Q_n(\alpha)}{n!(\lambda/\mu)n}$$

$$x \sum_{j=o}^{k-n-\nu-1} \binom{k-n-\nu-1}{j} \frac{n! H_j(\alpha)}{(n+j+1)!}$$

$$- \sum_{\nu=o}^{k-2} \frac{(-1)^{\nu} H_{\nu}(\alpha)}{\nu!(\lambda/\mu)^{\nu}} \sum_{k=o}^{k-\nu-2} \frac{Q_n(\alpha)}{n!(\lambda/\mu)^{n+1}}$$

$$x \sum_{j=o}^{k-n-\nu-2} \binom{k-n-\nu-2}{j} \frac{(n+1)! H_j(\alpha)}{(n+j+2)!} .$$

On constate que l'expression de $\Gamma_k^*(1, z)$, $(k > 1)$, contient encore une constante inconnue, à savoir $\Gamma_1^*(1, z)$. Celle-ci sera toutefois donnée par la formule suivante

$$\Gamma_1^*(1, z) = \frac{(\lambda_o + z) P_{10}^*(z)}{1 + \lambda_o P_{10}^*(z)} \quad ; \qquad (11)$$

valable pour tout modèle d'attente $M_n/M_n/1$ pour lequel le processus $\{X(t), t \geq 0\}$- état du système à l'instant t - est ergodique.
Notons que $P_{10}^*(z)$ est la transformée de Laplace de la probabilité

$$P_{10}(t) = \mathbb{P}\left[X(t) = 0 \mid X(0) = 1\right].$$

La démonstration de la formule (11) a été indiquée par HADIDI, [2], mais pour le Modèle A.
Une démonstration plus rigoureuse, basée sur la théorie de renouvellement, a été donnée par NATVIG, [5] , pour le Modèle C..
Cette démonstration de NATVIG se généralise aux autres modèles $M_n/M_n/1$ satisfaisant à l'hypothèse d'ergodicité [4].
L'expression de $P^*_{10}(z)$ dans le cas du Modèle B vaut, après quelques calculs:

$$P^*_{10}(z) = \frac{1}{\mu} \sum_{n=o}^{\infty} \frac{(n+1)\ (\lambda/\mu)^n}{(n-\alpha)\ (n-\alpha+1)} \quad ; \qquad (12)$$

$$\alpha = z/(\lambda-\mu) \qquad \text{et } 0 < \frac{\lambda}{\mu} < 1.$$

Les relations (11) et (12) permettent, enfin, d'éliminer la constante inconnue $\Gamma^*_1\ (1,\ z)$ de la formule (10). Dès lors celle-ci peut servir au calcul des moments de la V.A. T_k.
Calculons-en, par exemple, la moyenne :

$$m_1^{(k)} = E\ (T_k) = - \left. \frac{d\ \Gamma^*_k\ (1,\ z)}{dz} \right|_{z=o}$$

$$= \frac{1}{\mu-\lambda} \left\{1 - \sum_{j-1}^{k-1} \frac{(-1)^j}{j\ k} \binom{k}{j+1}\right\}, (\mu > \lambda) \ ; (13a)$$

$$= \frac{1}{\mu - \lambda} \sum_{r+1}^{1} \ ; \qquad (\mu > \lambda). \qquad (13b)$$

Ces résultats (13a) et(13b) s'obtiennent au prix de longs calculs, faut-il le dire. Enfin, le calcul des autres moments, à partir de (10), est une question de routine et de patience !

BIBLIOGRAPHIE.

1. CONOLLY, B.W. and CHAN, J. :

"Generalised birth and death queueing process : recent results". Adv. Appl. Prob. vol.9, (1977), pp. 125-140.

2. HADIDI, N. and CONOLLY, B.W. :

"On the reduction of congestion in queueing system". Statist. Research Report N°6, Institute of Mathematics, University of Oslo (1969).

3. HADIDI, N. :

"Busy period of queues with state dependent arrival and service rates". J.Appl. Prob., vol. 11, (1974), pp 842-848.

4. MANYA, N.:

"Modèles d'attente à interarrivées et durées de service dépendant de l'état du système". Thèse de doctorat en sciences, U.L.B. (1979).

5. NATVIG, B. :

" On a queueing model where potential customers are discouraged by queue length". Scand J. Statist., vol.2, (1975), pp 34-42.

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

FAMILLES DE GRAPHES D'INTERVALLES EMBOITES

Marc ROUBENS

Faculté Polytechnique de Mons
rue De Houdain 9
Belgium

ABSTRACT

We consider a familly of nested interval graphs and define conditions for this family to be lower-homogeneous, i.e. to present at least one numerical representation by intervals of the real line with origins independant from the index level.

1. *INTRODUCTION*

Considérons sur une droite, une famille finie {I(a),I(b),...} d'intervalles.

On appelle *graphe d'intervalles* le graphe dont les sommets a,b,..., représentent les intervalles, deux sommets étant joints si et seulement si les intervalles correspondants s'intersectent. Si A = {a,b,...} et si I représente l'ensemble des arêtes, on note le graphe d'intervalles G(A,I).

A une famille $\{G_i(A,I_i),\ i=1,\ldots,k\}$ de graphes d'intervalles correspondant des ensembles d'intervalles $\{I_i(a), a \in A\}$, i=1,...,k, de la droite tels que

$$(a,b) \in I_i \quad \text{ssi} \quad I_i(a) \cap I_i(b) \neq \phi$$

Chaque intervalle $I_i(a)$ peut être défini par une origine $g_i(a)$ et une longueur $q_i(a)$ telles que $I_i(a) = [g_i(a),\ g_i(a) + q_i(a)]$. Dès lors :

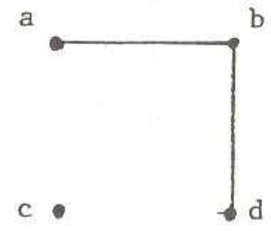
$$(a,b) \in I_i \leftrightarrow a\ I_i\ b \quad \text{ssi} \begin{cases} g_i(a) \leq g_i(b)+q_i(b) \\ g_i(b) \leq g_i(a)+q_i(a) \end{cases}$$

Une famille est composée de graphes d'intervalles $G_i(A,I_i)$ emboités ssi $I_1 \supseteq \ldots \supseteq I_k$

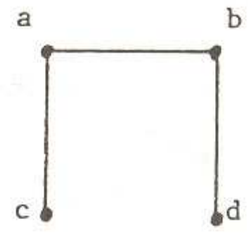
La famille des graphes d'intervalles emboités $G_i(A,I_i)$, i=1,...,k, est dite *inférieurement homogène* (lower-homogeneous family of nested interval graphs) si l'ensemble $\{I_i(a),\ i=1,\ldots,k, \forall\ a \in A\}$ est tel que

$$g_i(a) = g(a)\ ;\ i=1,\ldots,k,\ \text{tout } a \in A$$

Comme premier exemple, considérons

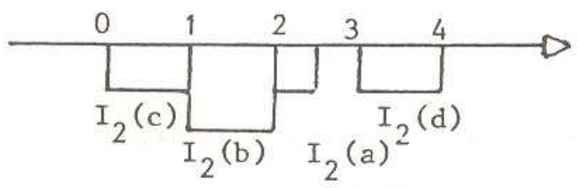


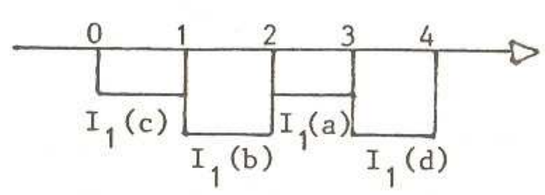
$G_2(A,I_2)$



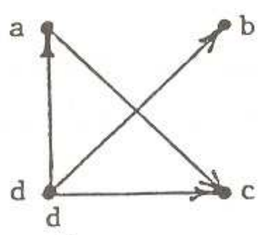
$G_1(A,I_1)$

On peut obtenir deux représentations par des intervalles dont les origines sont identiques pour chaque sommet de A :

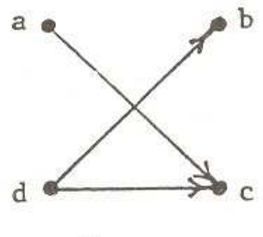




$\{G_i(A,I_i),\ i=1,2\}$ représente une famille inférieurement homogène de graphes d'intervalles emboités avec $I_2 \subset I_1$. A ces deux représentations par des intervalles on peut faire correspondre des *orientations transitives* des graphes complémentaires $\tilde{G}_i(A,P_i)$ telles que a P_i b ssi x>y, $\forall x \in I_i(a)$, $\forall y \in I_i(b)$ et $P_1 \subset P_2$

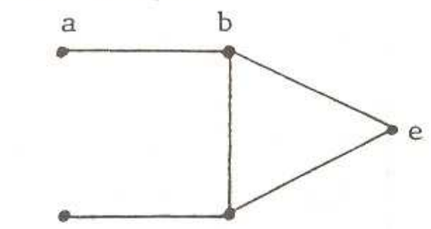


$\tilde{G}_2(A,P_2)$

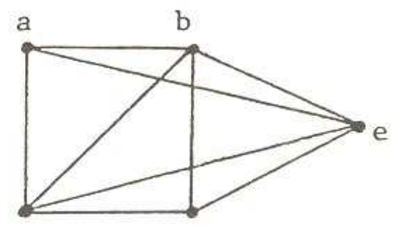


$\tilde{G}_1(A,P_1)$

Un second exemple

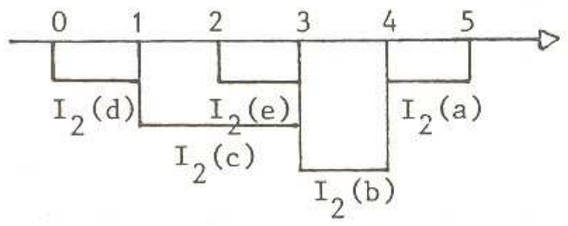


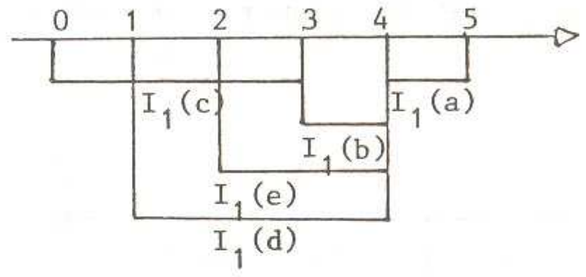
$G_2(A,I_2)$



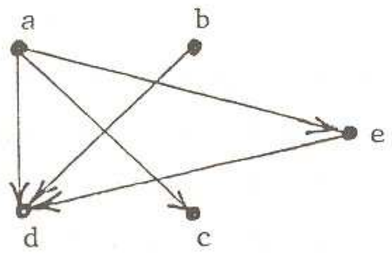
$G_1(A,I_1)$

permet de constater qu'une bi-représentation par des intervalles est encore possible :

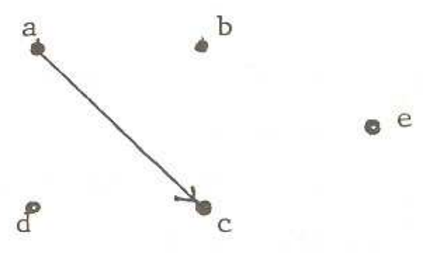




Des orientations transitives des graphes complémentaires $\tilde{G}_i(A,P_i)$ correspondant aux représentations par des intervalles donnent



$\tilde{G}_2(A,P_2)$



$\tilde{G}_1(A,P_1)$

Ici on a encore $P_2 \supset P_1$ mais la famille n'est pas inférieurement homogène ainsi qu'on le verra dans la suite car il est impossible d'obtenir une représentation de $G_i(A,I_i)$ telle que $g_i(a)=g(a)$, i=1,2.

Le théorème de Gilmore et Hoffman (1964) nous apprend qu'un graphe G(A,I) est un graphe d'intervalles si et seulement si G(A,I) est triangulé et son complémentaire $\tilde{G}(A,I)$ est transitivement orientable.

Un graphe d'intervalles est appelé graphe d'indifférence si, en vertu d'un résultat de Roberts (1969), il n'existe aucun sous-graphe partiel du type $K_{1,3}$ (graphe complet biparti avec deux sous-ensembles de sommets de cardinals 1 et 3). La terminologie est malheureusement inadéquate puisqu'un graphe d'intervalles représente également les indifférences relatives à une structure sous-jacente de préférence du type "ordre d'intervalles".
Pour déterminer si un graphe G(A,I) est un graphe d'intervalles, on contrôle l'absence de cycle de longueur 4 sans corde par l'algorithme de Tarjan (1976) et on recherche les cliques maximales par l'algorithme décrit dans l'ouvrage de Golumbic (1980) et inspiré d'un résultat de Fulkerson et Gross (1965).
Les cliques maximales doivent alors être orientées de manière que, lorsqu'un sommet apparaît dans plusieurs cliques maximales, celles-ci sont consécutives.

En se rappelant le résultat de Booth et Leuker (1976), ce problème se ramène à la recherche de matrices d'appartenance "cliques maximales-sommets" telles que chaque colonne indiquant (par 1) l'appartenance d'un sommet à une clique maximale présente une succession non interrompue de 1.

2. FAMILLE DE GRAPHES D'INTERVALLES EMBOITES INFERIEUREMENT HOMOGENE

ROUFENS et VINCKE (1984) ont introduit la notion de famille homogène en lignes d'ordres d'intervalles (row-homogeneous family of interval orders) : une famille d'ordres d'intervalles définis par les mises en ordre $(O_{L,i}, O_{C,i}, i=1,....,k)$ des lignes et colonnes des matrices d'adjacence (cf. définition dans Jacquet-Lagrèze (1978)) est homogène en ses lignes ssi $O_{L,i} = O_L$, $i=1,...,k$.

Ils ont montré que la condition nécessaire et suffisante pour obtenir une représentation numérique d'une famille d'ordres intervalles (A,P_i,I_i) telle que

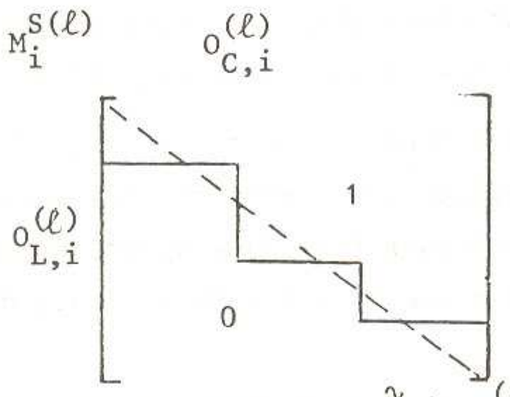
$$a\ P_i\ b \leftrightarrow g(a) > g(b)+q_i(b)$$

$$a\ I_i\ b \leftrightarrow \begin{cases} g(a) \leq g(b)+q_i(b) \\ g(b) \leq g(a)+q_i(a) \end{cases}$$

avec $I_{i-1} \supseteq I_i$, $P_i \subseteq P_i$, $0 \leq q_{i-1}(a) \leq q_i(a)$, $\quad a \in A$, $i=2,\ldots,k$

est que la famille emboitée d'ordres d'intervalles soit homogène en ses lignes

Considérons une famille de graphes d'intervalles $G_i(A,I_i)$. Les orientation transitives $\tilde{G}_i(A,P_i^{(\ell)})$, $\ell=1,2,3,\ldots$ (l'indice ℓ correspond aux différentes orientations possibles), associées à cette famille conduisent à des représentations matricielles étagées d'ordres d'intervalles induites par les couples $(\tilde{G}_i(A,P_i^{(\ell)}), G_i(A,I_i))$ telles que



Partant des orientations transitives $\tilde{G}_i(A,P_i^{(\ell)})$ associées aux graphes d'intervalles $G_i(A,I_i)$, $i=1,\ldots,k$, on obtient les mises en ordre $O_{L,i}^{(\ell)}$.

Il en résulte la proposition suivante :

Une famille de graphes d'intervalles emboités est inférieurement homogène si et seulement s'il existe des orientations transitives $(\ell_1,\ldots,\ell_k)$ telles que

$$O_{L,1}^{(\ell_1)} = O_{L,2}^{(\ell_2)} = \ldots = O_{L,k}^{(\ell_k)}$$

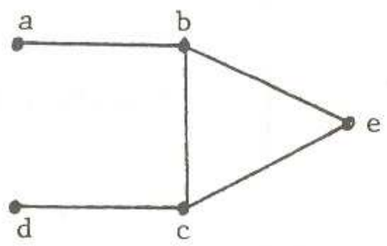
Il convient, afin d'obtenir des ordres totaux $O_{L,i}$, $O_{C,i}$, de travailler dans l'ensemble A/E_i où E_i représente une relation d'équivalence telle que

$$a\ E_i\ b \quad \text{ssi} \quad a\ I_i\ c \leftrightarrow b\ I_i\ c\ , \forall c \in A$$

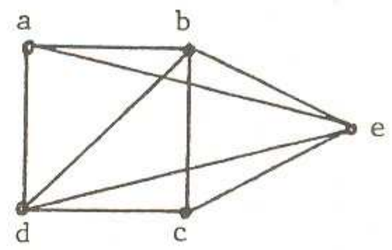
. *EXEMPLES*

3.1. Exemple 1

Reconsidérons l'exemple décrit dans la section 1



$G_2(A,I_2)$



$G_1(A,I_1)$ $\quad I_2 \subset I_1$

Les cliques maximales de $G_2(A,I_2)$ et $G(A,I_1)$ sont respectivement :

$$\begin{cases} C_1^{(2)} = \{a,b\} \\ C^{(2)} = \{b,c,e\} \\ C_3^{(2)} = \{c,d\} \end{cases} \qquad \begin{cases} C_1^{(1)} = \{a,b,d,e\} \\ C_2^{(1)} = \{b,c,d,e\} \end{cases}$$

Les mises en ordre des cliques correspondant à des matrices d'appartenance présentant des 1 consécutifs en colonne sont

$$\begin{cases} C_1^{(2)} > C_2^{(2)} > C_3^{(2)} \\ C_3^{(2)} > C_2^{(2)} > C_1^{(2)} \end{cases} \qquad \begin{cases} C_1^{(1)} > C_2^{(1)} \\ C_2^{(1)} > C_1^{(1)} \end{cases}$$

et tels que :

$$\begin{array}{c} \\ C_1^{(2)} \\ C_2^{(2)} \\ C_3^{(2)} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{bmatrix} 1 & 1 & & & \\ & 1 & 1 & & 1 \\ & & 1 & 1 & \end{bmatrix} \end{array} \qquad \begin{array}{c} \\ C_1^{(1)} \\ C_2^{(1)} \end{array} \begin{array}{c} \begin{array}{ccccc} a & b & c & d & e \end{array} \\ \begin{bmatrix} 1 & 1 & & 1 & 1 \\ & 1 & 1 & 1 & 1 \end{bmatrix} \end{array}$$

Il en découle les représentations matricielles (matrice d'adjacence) des ordres d'intervalles sous-jacents

$M_2^{S(1)}$ a b e c d

$O_{L,2}^{(1)}$:

	a	b	e	c	d
a					
b			1		
e					
c					
d			0		

$M_2^{S(2)}$ d c e b a

$O_{L,2}^{(2)}$: même matrice que $M_2^{S(1)}$

$M_1^{S(1)}$ a b e d c ($E_1^{(1)}$ = b e d)

$O_{L,1}^{(1)}$:

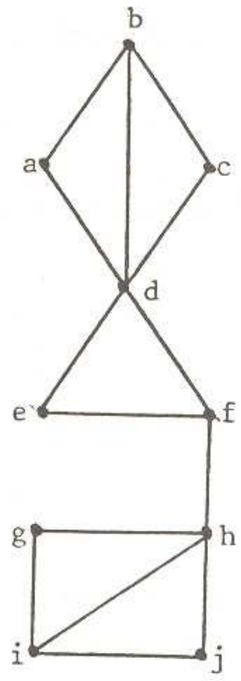
	a	b	e	d	c
a					
b			1		
e					
d					
c		0			

$M_1^{S(2)}$ c e d b a ($E_1^{(2)}$ = e d b)

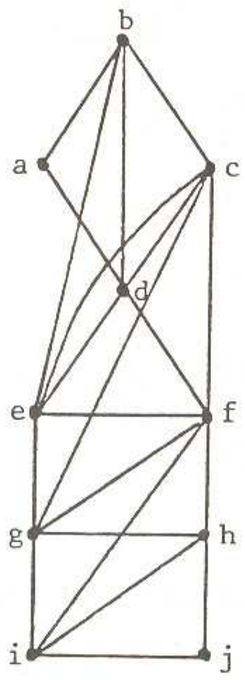
$O_{L,1}^{(2)}$: même matrice que $M_1^{S(1)}$

Aucun des deux ordres (a>b>e>c>d), (d>c>e>b>a) n'est compatible avec les préordres (a>b∿e∿d>c), (c>d∿e∿b>a); la famille $\{G_i(A,I_i)\}$ n'est pas inférieurement homogène. Notons que ni $G_2(A,I_2)$, ni $G_1(A,I_1)$ ne comportent de sous-graphes partiels du type $K_{1,3}$. Dès lors, les ordres d'intervalles sous-jacents aux orientations transitives sont des quasi-ordres et les graphes d'intervalles sont des graphes d'indifférence.

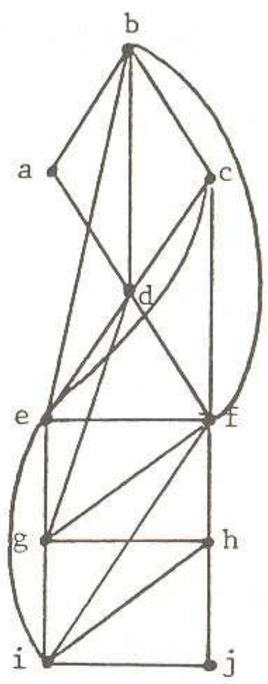
3.2. Exemple 2



$G_3(A,I_3)$

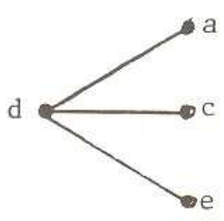


$G_2(A,I_2)$



$G_1(A,I_1)$

Aucun de ces trois graphes n'est un graphe d'indifférence car le sous-graphe partiel du type $K_{1,3}$



est présent dans $G_i(A,I_i)$, i=1,2,3.

Notons encore que dans $G_2(A,I_2)$, (h,i) forme une classe d'équivalence - ils ont même comportement à l'égard de tous les autres sommets - et que (h,i) et (e,f) sont des classes d'équivalence de $G_1(A,I_1)$.

Les cliques maximales de $G_3(A,I_3)$, $G_2(A,I_2)$ et $G_a(A,I_1)$ sont respectivement :

$$\begin{cases} C_1^{(3)} : \{a,b,d\} \\ C_2^{(3)} : \{b,c,d\} \\ C_3^{(3)} : \{d,e,f\} \\ C_4^{(3)} : \{f,h\} \\ C_5^{(3)} : \{g,h,i\} \\ C_6^{(3)} : \{h,i,j\} \end{cases} \qquad \begin{cases} C_1^{(2)} : \{a,b,d\} \\ C_2^{(2)} : \{b,c,d,e\} \\ C_3^{(2)} : \{c,d,e,f\} \\ C_4^{(2)} \quad \{d,e,f,g\} \\ C_5^{(2)} \quad \{f,g,h,i\} \\ C_6^{(2)} \quad \{h,i,j\} \end{cases} \qquad \begin{cases} C_1^{(1)} : \{a,b,d\} \\ C_2^{(1)} : \{b,c,d,e,f\} \\ C_3^{(1)} : \{d,e,f,g,h,i\} \\ C_4^{(1)} : \{h,i,j\} \end{cases}$$

Les mises en ordre des cliques correspondant à des matrices d'appartenance qui présentent des 1 consécutifs en colonnes sont :

$$\begin{cases} C_1^{(3)} > C_2^{(3)} > C_3^{(3)} > C_4^{(3)} > C_5^{(3)} > C_6^{(3)} \\ C_2^{(3)} > C_1^{(3)} > C_3^{(3)} > C_4^{(3)} > C_5^{(3)} > C_6^{(3)} \\ C_1^{(3)} > C_2^{(3)} > C_3^{(3)} > C_4^{(3)} > C_6^{(3)} > C_5^{(3)} \\ C_2^{(3)} > C_1^{(3)} > C_3^{(3)} > C_4^{(3)} > C_6^{(3)} > C_5^{(3)} \end{cases}$$

ainsi que les classements inverses,

$$C_1^{(2)} > C_2^{(2)} > C_3^{(2)} > C_4^{(2)} > C_5^{(2)} > C_6^{(2)}$$

et le classement inverse,

$$C_1^{(1)} > C_2^{(1)} > C_3^{(1)} > C_4^{(1)}$$

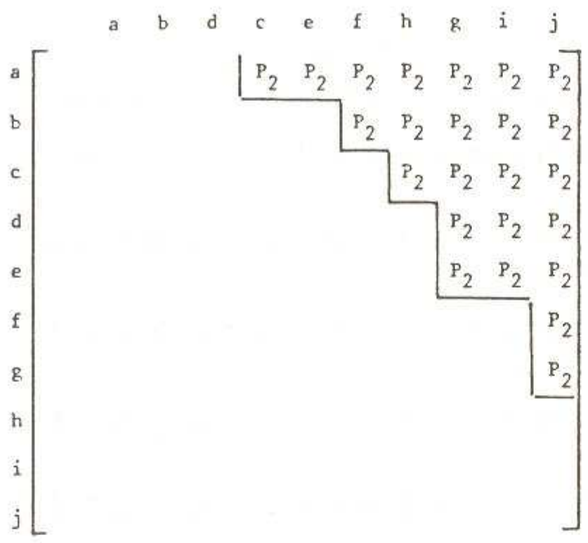
et le classement inverse.

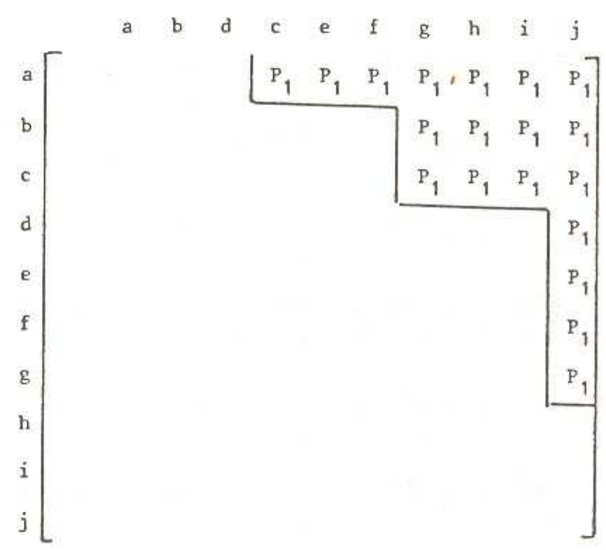
Il en découle les mises en ordre selon les lignes des différentes représentations matricielles étagées des ordres d'intervalles sous-jacents :

Mise en ordre des cliques maximales	O_L
$C_1^{(3)} > C_2^{(3)} > C_3^{(3)} > C_4^{(3)} > C_5^{(3)} > C_6^{(3)}$	a > b > c > d > e > f > g > h > i
$C_2^{(3)} > C_1^{(3)} > C_3^{(3)} > C_4^{(3)} > C_5^{(3)} > C_6^{(3)}$	c > b > a > d > e > f > g > h > i
$C_1^{(3)} > C_2^{(3)} > C_3^{(3)} > C_4^{(3)} > C_6^{(3)} > C_5^{(3)}$	a > b > c > d > e > i > g > h > f
$C_2^{(3)} > C_1^{(3)} > C_3^{(3)} > C_4^{(3)} > C_6^{(3)} > C_5^{(3)}$	a > b > a > d > e > i > g > h > f
$C_1^{(2)} > C_2^{(2)} > C_3^{(2)} > C_4^{(2)} > C_5^{(2)} > C_6^{(2)}$	a > b > c > d > e > f > g > h ∿ i > j
$C_1^{(1)} > C_2^{(1)} > C_3^{(1)} > C_4^{(1)}$	a > b > c > d > e ∿ f > g > h ∿ i > j

Les ordres {a > b > c > d > e > f > g > h > i} et {i > h > g > f > e > d > c > b > a} sont donc compatibles avec la famille $\{G_i(A,I_i)\}$. Le premier de ces deux ordres donne les représentations matricielles des ordres d'intervalles sous-jacents à $\{G_i(A, I_i)\}$.

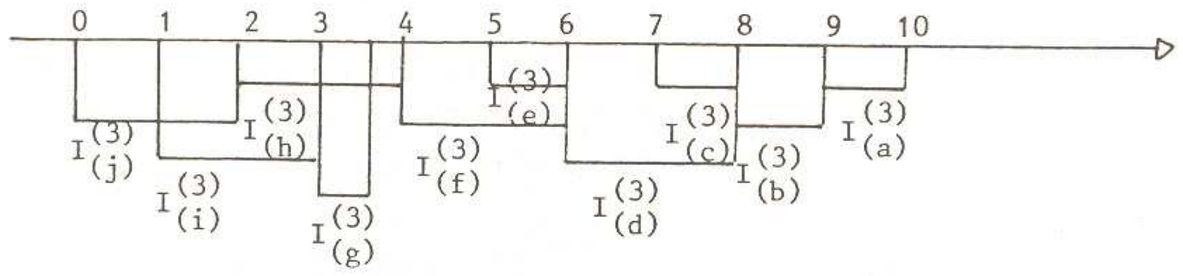
	a	b	d	c	e	f	h	g	i	j
a		P_3	P_3	P_3	P_3	P_3	P_3	P_3		
b			P_3	P_3	P_3	P_3	P_3	P_3		
c			P_3	P_3	P_3	P_3	P_3	P_3		
d					P_3	P_3	P_3	P_3		
e					P_3	P_3	P_3	P_3		
f						P_3	P_3	P_3		
g								P_3		
h										
i										
j										

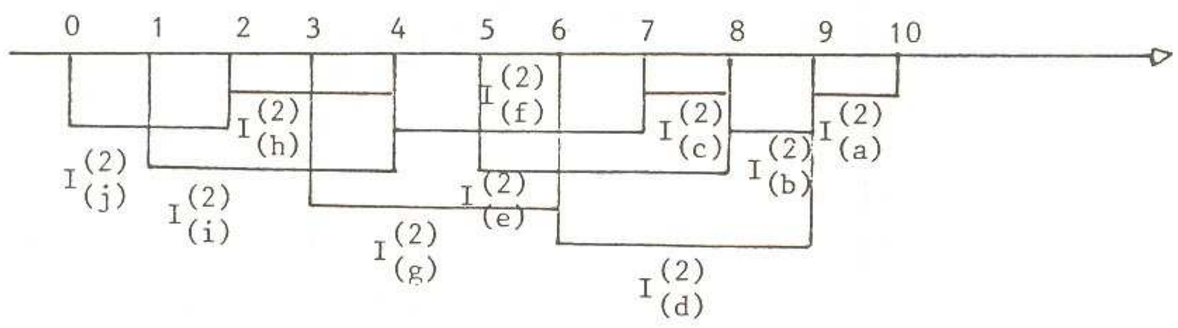


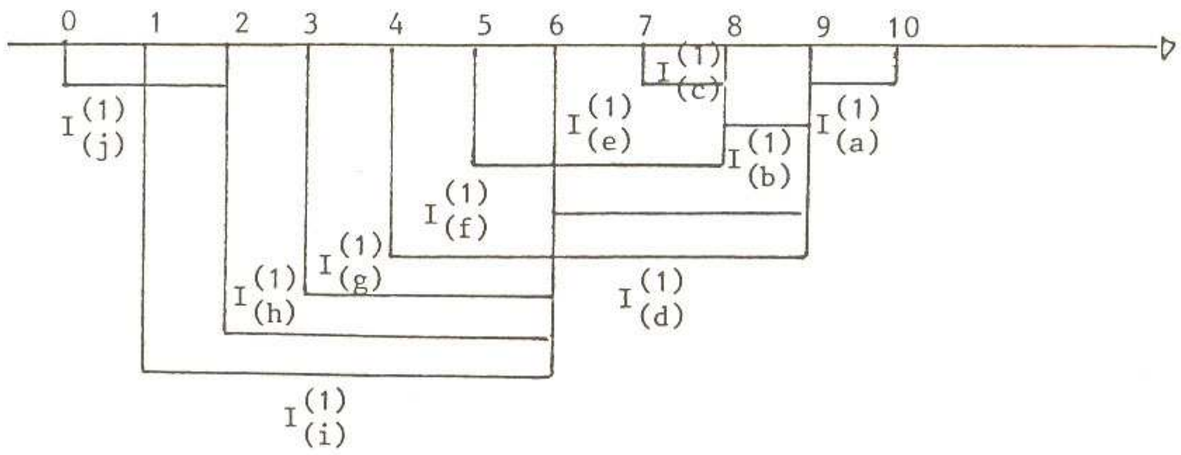


On constatera que les ordres 0_C ne coïncident pas (ordres d'intervalles !).

Les représentations avec origines communes sont les suivantes :







4. *BIBLIOGRAPHIE*

BOOTH K.S., LEUKER G.S.,
Testing for the consecutive ones property, interval graphs, and graph planarity using PQ-tree algorithms,
J. Comput. Syst. Sci. 13, 335-379 (1976).

FULKERSON D.R., GROSS O.A.
Incidence matrices and interval graphs,
Pacific J. Math. 15, 835-855 (1965).

GILMORE P.C., HOFFMAN A.J.
A characterization of comparability graphs and of interval graphs,
Candad.J. Math.16,539-548 (1964).

JACQUET-LAGREZE E.
Représentation de quasi-ordres et de relations probabilistes transitives sous forme standard et méthodes d'approximation,
Math. Sci. hum. 63, 5-24 (1978).

GOLUMBIC M.C.
Algorithmic graph Theory and Perfect Graphs
Academic Press, Inc. New York (1980).

ROBERTS F.S.,
Indifference Graphs, in "Proof Techniques in Graph Theory" (F. Harary, ed.)
pp.139-146, Academic Press, Inc. New York (1969).

ROUBENS M., VINCKE Ph.,
On Families of Semiorders and interval Orders Imbedded in a Valued Structure of Preference : a survey,
Information Sciences (1984).

TARJAN R.E.
Maximum cardinality search and chordal graphs.
Standford Univ. Unpublished Lecture Notes CS 259 (1976).

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

RECURRENCE OU PERIODICITE

R. SNEYERS

Institut royal météorologique de Belgique
Avenue Circulaire 3
1180 Bruxelles
Belgique

ABSTRACT

The theory of autoregressive series and of random periodic series is briefly recalled. It is shown that the series of sample autocovariances enables to decide of the type of model to adjust to the original series. Two examples are considered: the series of daily averages of the athmospheric pressure at Uccle and the series of annual Wolf sunspot numbers.

1. *Introduction.*

L'une des applications les plus importantes de la statistique mathématique à la géophysique est la représentation des séries chronologiques d'observations au moyen de modèles stochastiques en vue de prévisions. Parmi ceux-ci deux types de modèles jouent un rôle majeur : les modèles récurrents ou autorégressifs et les modèles périodiques. Dans le premier cas, la série chronologique obéit à une équation récurrente non homogène dont le terme indépendant est une quantité aléatoire simple; dans le second, à une quantité aléatoire près, la série se représente à l'aide de la somme de fonctions sinusoïdales de périodes déterminées. Il va de soi que les deux modèles se distinguent nettement en ce qui concerne leur pouvoir de prédiction, le modèle récurrent n'autorisant que la prévision à courte échéance et le modèle périodique permettant au contraire la prévision à longue échéance.

L'objet de la présente note est de rappeler brièvement les propriétés des deux types de modèles et les moyens de les distinguer; de plus, deux exemples sont donnés qui illustrent l'un et l'autre modèle.

2. *Les séries aléatoires récurrentes.*

Une série chronologique $x_1, x_2, \ldots, x_n$ est dite aléatoire récurrente (autorégressive) si elle vérifie une équation de la forme :

$$x_{i+k} = a_0 + a_1 x_i + a_2 x_{i+1} + \ldots + a_k x_{i+k-1} + e_{i+k} \qquad (1)$$

où $a_0, a_1, \ldots, a_k$ sont des constantes et où e_{i+k} sont des quantités aléatoires simples de moyennes nulles et de même répartition. De plus, à la condition que $a_1 \neq 0$, elle est dite d'ordre k.

Avec var $e_i = \sigma^2$ on a aussi : $\operatorname{cov}(x_i, e_j) = 0$ pour $i < j$ et $\operatorname{cov}(x_i, e_i) = \sigma^2$.

Toute solution de (1) peut se mettre sous la forme de la somme de solutions particulières de l'équation homogène :

$$x_{i+k} = a_1 x_i + a_2 x_{i+1} + \ldots + a_k x_{i+k-1} \qquad (2)$$

dont la solution générale est de la forme :

$$\Sigma \lambda_j x_i^{(j)} \text{ avec } x_i^{(j)} = b_r^i,\ b_s^i \sin \alpha_s i,\ b_s^i \cos \alpha_s i$$

sachant que b_r et b_s $(\cos \alpha_s + \sqrt{-1} \sin \alpha_s)$ sont les racines réelles ou complexes de l'équation caractéristique associée :

$$z^k - a_k z^{k-1} - a_{k-1} z^{k-2} \ldots - a_1 = 0 \qquad (3)$$

On en déduit que si x_i est une série stationnaire, on doit avoir :

$|b_r|<1$ et $|b_s|<1$, donc aussi $|a_1|<1$.

De plus, la série x_i est la somme de séries amorties de façon continue ou oscillante. Dans le second cas, on peut la considérer comme "pseudo-périodique".

Par ailleurs, si on pose $v_k = \mathrm{cov}(x_i, x_{i+k})$ pour $k>0$, on établit que la série des autocovariances v_k vérifie l'équation homogène (2). Il s'ensuit que $\lim v_k = 0$ lorsque k tend vers l'infini.

Enfin, si on désigne par $\overline{v}_k$ les autocovariances empiriques de la série, c'est-à-dire :

$$\overline{v}_k = [\Sigma x_i x_{i+k} - (\Sigma x_i)(\Sigma x_{i+k})/(n-k)]/(n-k-1) \qquad (4)$$

on montre que la série des $\overline{v}_k$ vérifie une équation du type (1) où toutefois la quantité e_{i+k} n'est pas aléatoire simple. De ce fait, la série des $\overline{v}_k$ ne présente qu'approximativement les caractères pseudo-périodiques de la série des v_k mais apparaît habituellement comme une série persistante (non aléatoire simple).

3. *Les séries aléatoires périodiques.*

Une série chronologique $x_1, x_2, \ldots, x_n$ est dite aléatoire périodique si elle peut se mettre sous la forme :

$$x_i = a_0 + \sum_j (a_j \sin \alpha_j i + b_j \cos \alpha_j i) + e_i \qquad (1)$$

où a_0, a_j, b_j et α_j sont des constantes et où e_i sont des quantités aléatoires simples de moyennes nulles et de même répartition.

Avec $c_j = \sqrt{a_j^2 + b_j^2}$ et $b_j/a_j = \mathrm{tg}\,\varphi_j$, (1) peut encore s'écrire :

$$x_i = a_0 + \Sigma c_j \sin(\alpha_j i + \varphi_j) + e_i \qquad (2)$$

Avec $\mathrm{var}\, e_i = \sigma^2$ on en tire immédiatement :

$$E(x_i) = a_0 \qquad v_0 = \mathrm{var}\, x_i = \Sigma c_j^2/2 + \sigma^2$$

$$v_k = \mathrm{cov}(x_i, x_{i+k}) = \Sigma (c_j^2/2) \cos \alpha_j k \qquad (3)$$

d'où il résulte que la série des autocovariances d'une série aléatoire périodique est une série périodique.

Si l'on définit la variance empirique $\overline{v}_0$ et les autocovariances empirique $\overline{v}_k$ à l'aide de la formule 2(4), il vient : $E(\overline{v}_0) \cong v_0$, $E(\overline{v}_k) \cong v_k$. De plus, si e_i est une variable normale, on a aussi :

$$\text{var } \overline{v}_0 = 2\sigma^4/(n\text{-}1) \quad , \quad \text{var } \overline{v}_k = \sigma^4/(n\text{-}k\text{-}1) \quad \text{et} \quad \text{cov}(\overline{v}_k , \overline{v}_l) = 0 \text{ pour } k \neq l > 0. \tag{4}$$

On en déduit que la série $\overline{v}_k$ est également une série aléatoire périodique ayant les mêmes périodes que la série originale, la variance de $\overline{v}_k$ étant toutefois croissante.

En particulier, si on pose $r_j^2 = \sigma^2/c_j^2$, il est clair que la périodicité apparaîtra d'autant mieux dans la série que r_j^2 est petit. Par ailleurs, le même rapport calculé pour la série $\overline{v}_k$ devient en vertu de (4) :

$$r_j^2(\overline{v}_k) = [\,\sigma^4/(n\text{-}k\text{-}1)]/(c_j^4/4) = \frac{4\,r_j^2}{n\text{-}k\text{-}1} \cdot r_j^2 \tag{5}$$

ce qui montre que $r_j^2(\overline{v}_k) < r_j^2$ dès que $(n\text{-}k\text{-}1) > 4\,r_j^2$.

Il s'ensuit que la période apparaîtra mieux dans la série $\overline{v}_k$ dès que n est suffisamment grand.

Le calcul des autocovariances empiriques est donc un excellent moyen de détection des périodicités dans les séries chronologiques.

4. *Estimation de la récurrence ou de la composante périodique.*

L'estimation des constantes peut se faire dans les deux cas par la méthode des moindres carrés.

Utilisant les notations matricielles, sachant que θ' est le transposé de θ, on pose dans le cas de la récurrence :

$$x = \| x_{i+k} \| \quad , \quad u = \|1\ x_i\, x_{i+1} \cdots x_{i+k\text{-}1} \| \quad , \quad \theta' = \| a_0\ a_1 \ldots a_k \| , \; e = \| e_{i+k} \|$$

tandis que dans le cas de la série aléatoire périodique, on fait :

$$x = \| x_i \| \quad , \quad u = \| 1 \sin \alpha_1 i \cos \alpha_1 i \sin \alpha_2 i \cos \alpha_2 i \ldots \| \quad , \quad \theta' = \|a_0\ a_1\ b_1\ a_2\ b_2 \cdots \| ,$$

$$e = \| e_i \|$$

d'où il résulte que les équations 2(1) et 3(1) se mettent sous la forme commune :

$$x = u\theta + e \tag{1}$$

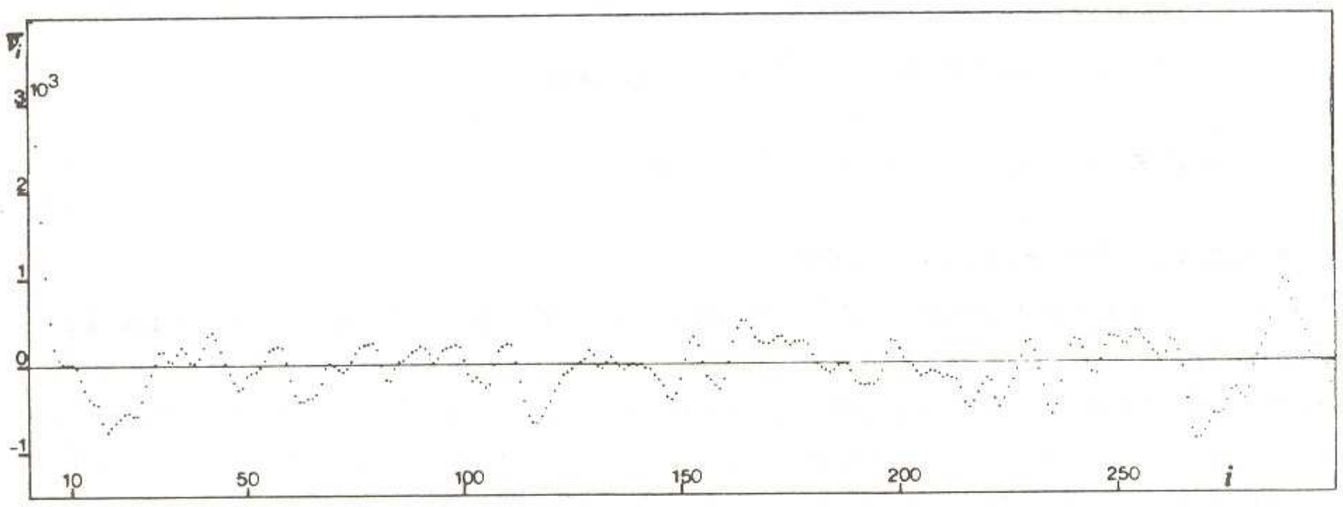


FIG. 1. — Autocovariances empiriques $\overline{v}_i$ pour $i = 1.2,...,300$, de la série des moyennes journalières de la pression atmosphérique à Uccle (d'après deux années d'observations).

L'estimation de θ s'obtient en minimisant la forme quadratique :

$$e'e = (x - u\theta)'(x - u\theta) \qquad (2)$$

ce qui donne l'estimation :

$$\hat{\theta} = (u'u)^{-1} u'x \quad \text{et} \quad \text{var}\,\hat{\theta} = (u'u)^{-1}\sigma^2 \qquad (3)$$

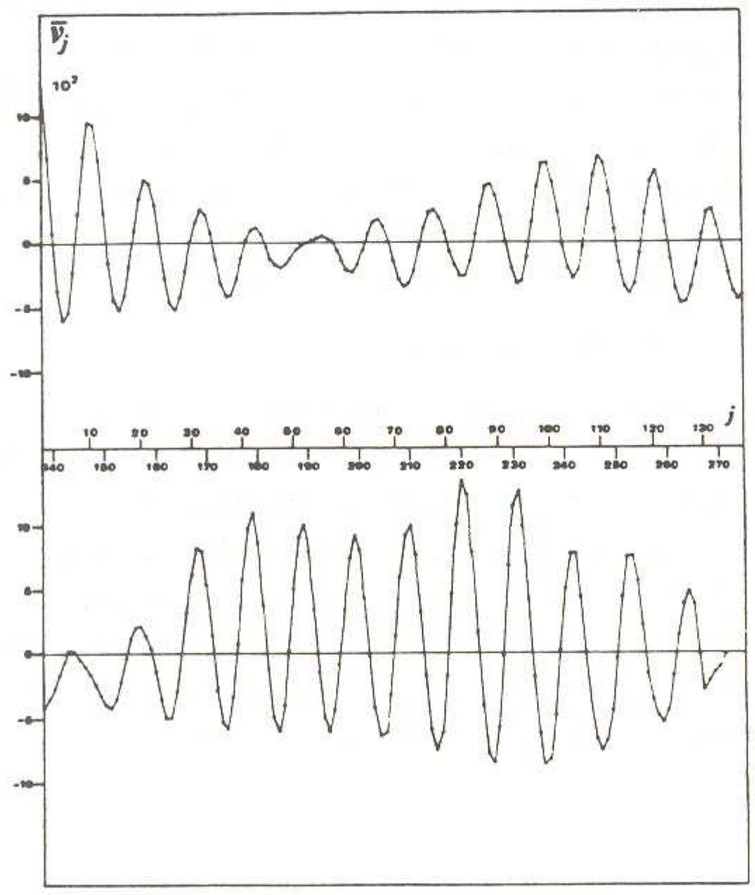


Fig.2 — Autocovariances empiriques $\overline{v}_j$ pour $j = 1, 2, ..., 271$ de la série des nombres de Wolf de 1700 à 1972.

De plus, on peut estimer σ^2 à l'aide de la relation :

$$\hat{\sigma}^2 = e'e/(n-n_1)$$

où n_1 est le nombre de paramètres estimés.

Dans le cas d'une série aléatoire récurrente, l'estimation peut se faire en ajustant successivement des récurrences d'ordre croissant 1, 2, ..., k jusqu'à l'obtention d'un résidu aléatoire simple. Pour une série aléatoire périodique, une première estimation des α_j peut être déduite du graphique des autocovariances empiriques $\overline{v}_k$, l'estimation finale étant ensuite obtenue en retenant celle qui donne la plus grande valeur à c_j dans un intervalle $(\alpha_{j_1}, \alpha_{j_2})$ entourant la première estimation de α_j .

5. *Exemples.*

a) La série des moyennes journalières de la pression atmosphérique à Uccle (deux années d'observations).

La figure 1 donne la représentation graphique des autocovariances empiriques $\overline{v}_k$ pour $k = 1, 2, ..., 300$.

Il apparaît que la série est persistante, mais ne présente aucune périodicité systématique. Des ajustements successifs montrent que la série de pressions obéit à une récurrence aléatoire d'ordre 2. La pseudo période des solutions de l'équation récurrente homogène correspondante est 16,7 j avec une erreur type d'estimation de 3,4 j. Cette pseudo période est en bon accord avec le phénomène de vacillation que l'on observe dans l'atmosphère. De plus, le modèle récurrent a permis de faire une bonne prévision statistique des extrêmes mensuels des moyennes journalières de la pression, ceux observés durant une période de trente années se révélant en bon accord avec la répartition théorique.

b) Le nombre de Wolf de taches solaires.

La figure 2 donne les autocovariances empiriques $\overline{v}_k$ des nombres de Wolf moyens annuels observés de 1700 à 1972 pour $k = 1, 2, ..., 271$. Le graphique révèle l'existence d'une périodicité de l'ordre de 10,5 ans dont l'amplitude varie elle-même avec une période de l'ordre de 200 ans. Des essais successifs par moindres carrés du modèle qu'on en déduit permet de trouver les estimations $\hat{T} = 10{,}494$ et $\hat{T}' = 200{,}5$. Le phénomène des taches solaires apparaît donc comme dû principalement à un phénomène de battement entre deux ondes de périodes voisines de 10,5 ans. Le résidu des périodicités estimées a été traité de la même manière et en continuant de la sorte d'autres périodicités ont été tirées de la série.

Au total 26 fonctions périodiques de la forme $a_j \sin \alpha_j i + b_j \cos \alpha_j i$ ont été retirées avant que la sélection n'ait été arrêtée par un test de signification qui ne sera pas détaillé ici.

Une comparaison a été faite entre la prédiction des nombres de Wolf qu'on peut en déduire et les nombres observés de 1973 à 1983. La décroissance du nombre de Wolf jusqu'en 1976 a été prévue de façon acceptable de même qu'un accroissement de ce nombre au delà, mais l'écart entre les valeurs maximales observées et la prédiction est trop grande pour rendre le modèle admissible. En réalité, la prédiction fournie par un modèle formé des premières sinusoïdes suggérées par la série des autocovariances et un résidu autorégressif du second ordre apparaît par contre comme acceptable. Le phénomène serait donc du à la combinaison d'un effet périodique et d'un effet récurrent.

Références.

Anderson, T.W., 1971 : The Statistical Analysis of Time Series, Wiley

Kendall, M.G., and Stuart, A., 1963 : The Advanced Theory of Statistics, 3 vols, Griffin.

Kendall, Sir Maurice, 1975 : Time-Series, Griffin.

Sneyers, R., 1975 : Sur l'analyse statistique des séries d'observations, O.M.M. Note technique n° 143.

Sneyers, R., 1975 : Les séries aléatoires récurrentes, Application aux moyennes journalières de la pression à Uccle, in Hommage à-Hulde aan Jacques Van Mieghem, Inst. R. Mét. Pub. A, n° 91.

Sneyers, R., 1976 : Application of Least Squares to the Search for Periodicities, J. Appl. Met., Vol. 15, No 4, pp. 387-393.

Sneyers, R. et P. Cugnon , 1984 : On the predictability of the Wolf sunspot number, European Geophysical Society, 10th Annual Meeting, Louvain la Neuve, 30 July- 3 August 1984.

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

ON AN OPTIMAL POLICY FOR DIVERTING TRAFFIC FLOW FROM A CONGESTED AREA

TRÂN-QUÔC-TÊ

ABSTRACT

This work investigates policies for diverting traffic flow from a main way where some flow-stopping incident has occurred. The model chosen for describing the main way congestion is basically a queuing model: vehicles that are trapped by the accident can leave the jam, but at a slower than normal rate, and the congestion will terminate when the waiting queue becomes empty.

The new feature introduced is that there exists a branching point in the upstream of the congested area that gives a controller (either human or automatic) the ability of diverting a fraction of vehicular flow towards some uncongestioned auxilary way. The objective aimed is to minimize a cost function that measures 1) the amplitude of the congestion as the total number of vehicles involved in the jam, the jam duration, and the total vehicle-hours waited, and 2) diversion costs that may take into account the lengthening in travel time incurred by diverted drivers.

Traffic diversion policies are analyzed by using a Markov (birth-and-death) model. It is shown that the best rule leads simply to divert an arriving vehicle if and only if the current queue length exceeds some given upper limit.

INTRODUCTION

Decisional models for vehicular traffic generally deal with routing a traffic flow (viewed as either discrete or continous) in a network [01, 02, 03,...], or with synchronizing a series of traffic lights [04, 05, 06,...]. These models usually consider normal (i.e. expected) conditions on traffic flow, and analyze steady-state behavior of traffic systems.

We examine in this work another kind of situations which, while "accidental", may be nevertheless of some interest : that concerns rules for diverting a vehicular flow from a main way where some flow-stopping incident has occurred. Among queuing models that describe the resulting jam, we retain that of Gaver[07]: as far as traffic is concerned, - and only as far as traffic is concerned -, a car accident mainly results in a physical obstacle; the latter limits traffic flow to a slower than normal rate, because jammed cars get in each others way. The congestion is considered as completely dissipated when the jam queue length falls below some non-congestion level. Traffic can then flow freely again.

We introduce a decisional aspect to the problem by assuming that this stoppage happens in the context described by map (figure) 1.

Whenever a car arrives at the branching point, one has to decide whether this car should be diverted or not. Such a dichotomic choice, that should use control informations about the current "state" of the system, must take into account the following two bjectives : 1) reduction of ineffectiveness of the congestion, .nd 2) reduction of diversion costs.

As measures of the first kind, we consider :

- the jam duration, D,
- the total number of trapped (i.e. non-diverted) cars, N,
- the total vehicle-hours waited, that is the total waiting delay of trapped cars, W.

Concerning diversion costs, we assume that any diverted driver incurs the same lengthening in his travel time, hence a penalty proportional to the total number M of diverted cars.

Optimal diversion policies are those minimizing expectation of

$$C = \gamma_d D + \gamma_w W + \gamma_n N + \gamma_m M, \qquad (1)$$

where the γ's are given non-negative numbers.

A STOCHASTIC MODEL FOR CONTROLLING THE JAM DISSIPATION

Let n_1 denote the number of cars present at the jam at the beginning of the congestion, and n_2 the non-saturation level below which traffic can flow freely and the jam is completely dissipated.

We assume that during this phase vehicles arrive at the intersection according to a time-stationary Poisson process of rate λ, and that trapped vehicles leave the jam according to a "death" process of rates $\{\mu_i\}$; that is, if the jam queue length at some time t is i ($i \geqslant n_2$), then, independently of the "past history" :

$$\Pr[\,0 \text{ departure during } [t,t+\Delta t]\,] = 1 - \mu_i \Delta t + o(\Delta t),$$

$$\Pr[\,1 \text{ departure during } [t,t+\Delta t]\,] = \mu_i \Delta t + o(\Delta t),$$

$$\Pr[\,> 1 \text{ depart. during } [t,t+\Delta t]\,] = o(\Delta t),$$

for any $\Delta t \geqslant 0$. Allowing the departure process to be state-dependent, we can describe situations in which trapped cars get in each others way, by assuming for example that the sequence $\{\mu_i / i\}$ decreases to 0 as $i \to \infty$.

Since arrival and departure processes are markovian and time-stationary, we restrict ourselves to diversion policies specified by :

$$\delta = \{\delta_i ;\ i \geqslant n_2\}, \tag{2}$$

such that :

$$\delta_i \in [0,1],\ \forall\ i. \tag{3}$$

Use of policy δ means that, whenever a vehicle arrives at the intersection while i other ones are present in the jam queue, then that vehicle is diverted with probability δ_i (and non-diverted with probability $1 - \delta_i$).

For $n_1 \geqslant n_2$, the congestion may be considered as a period or the first passage of the jam queue length from n_1 to $n_2 - 1$. Hence, the jam is decomposable into a juxtaposition of $n_1 - n_2 + 1$ periods, <u>period i</u> ($i = n_2, \ldots, n_1$) being that for the first passage of the queue from i to i - 1 cars, see Figure 2.

Assume a given diversion policy is being used. We denote (i) the duration of a period i, (ii) the total time loss during this period, (iii) the number of cars that join the jam during this period, and (iv) the number of cars diverted during this period, by D_i, W_i, N_i, and M_i respectively. The total cost incurred during the decongestion phase is the sum of costs related to each of the above-mentioned periods :

$$C = \sum_{i=n_2}^{n_1} C_i, \text{ where} \tag{4}$$

$$C_i = \gamma_d D_i + \gamma_w W_i + \gamma_n N_i + \gamma_m M_i \tag{5}$$

are statistically independent of each other, by our markovian assumptions. We first derive a recursive formula for the joint Laplace transforms

$$\Phi_i(d,w,x,y) = \mathcal{E}\left[e^{-dD_i} . e^{-wW_i} . x^{N_i} . y^{M_i} \right], \tag{6}$$

$$i \geqslant n_2, \quad d,w \geqslant 0, \quad |x|, |y| \leqslant 1,$$

before investigating properties of optimal policies.

Consider now the evolution of the jam queue at some time epoch. The next "event" that will occur is either a departure from the jam or an arrival to the intersection (in the later case, the arriving car may be diverted or not). Hence, for $i \geqslant n_2$:

$$(D_i,W_i,N_i,M_i)$$

$$\doteq \begin{cases} (T_i, iT_i, 0, 0), & \text{with probability } \mu_i/(\lambda+\mu_i), \\ (T_i', iT_i', 0, 1) + (D_i', W_i', N_i', M_i'), & \text{with probability } \delta_i\lambda/(\lambda+\mu_i), \\ (T_i'', iT_i'', 1, 0) + (D_{i+1}'', W_{i+1}'', N_{i+1}'', M_{i+1}'') + (D_i'', W_i'', N_i'', M_i''), & \\ \qquad \text{with probability } (1-\delta_i)\lambda/(\lambda+\mu_i), & \end{cases}$$

where the symbol $\doteq$ means equality of distributions, and where, by our markovian assumptions :

- $T_i \doteq T_i' \doteq T_i'' \doteq$ a random variable distributed exponentially with mean $(\lambda+\mu_i)^{-1}$,
- $(D_i', W_i', N_i', M_i') \doteq (D_i, W_i, N_i, M_i)$, and is statistically independent of T_i',
- $(D_{i+1}'', W_{i+1}'', N_{i+1}'', M_{i+1}'') \doteq (D_{i+1}, W_{i+1}, N_{i+1}, M_{i+1})$, and is independent of T_i'' and of $(D_i'', W_i'', N_i'', M_i'') \doteq (D_i, W_i, N_i, M_i)$.

Hence, using the fact that the Laplace transform of (the distribution function of) T_i is :

$$\mathcal{E}[e^{-tT_i}] = (\lambda+\mu_i)/(t+\lambda+\mu_i), \quad t \geqslant 0,$$

the above inductive decomposition of (D_i,W_i,N_i,M_i) leads readily to:

$$\Phi_i(d,\omega,x,y) = \frac{\mu_i + \delta_i\lambda y\Phi_i(d,\omega,x,y) + (1-\delta_i)\lambda x\,\Phi_{i+1}(d,\omega,x,y)\Phi_i(d,\omega,x,y)}{d + i\omega + \lambda + \mu_i} \qquad (7)$$

This formula links Φ_i and Φ_{i+1}, and allows recursive computation of Φ_i, $\forall$ i, provided some Φ_j is known :

$$\ldots \leftarrow \Phi_{j-2} \leftarrow \Phi_{j-1} \leftarrow \Phi_j \rightarrow \Phi_{j+1} \rightarrow \Phi_{j+2} \rightarrow \ldots$$

It will be especially so if some $\delta_j = 1$, that is, if we decide to divert any vehicle that arrives while the jam size is currently j. In this case, as one can easily anticipate, Φ_j no longer depends on Φ_{j+1}, and the above formula reduces to a linear equation in Φ_j whose solution is :

$$\Phi_j^1(d,\omega,x,y) = \mu_j / (d + j\omega + \lambda + \mu_j - \lambda y). \qquad (8)$$

Note that, from a practical point of view, there generally exists a constraint of capacity on the main way that implies automatic diversion whenever the jam size reaches some oversaturation level. We now show that, even when such a constraint is not imposed, the best diversion rule merely consists in diverting an arriving car if and only if the current queue length reaches some *diversion level*.

Recall that optimal diversion policies are those minimizing(+) :

$$\bar{C} = \sum_{i=n_1}^{n_2} \bar{C}_i . \qquad (9)$$

Taking partial derivatives of Φ_i, we obtain from (7) :

$$\begin{aligned} \bar{D}_i &= \{1 + (1-\delta_i)\lambda \bar{D}_{i+1}\} / \mu_i , \\ \bar{W}_i &= \{i + (1-\delta_i)\lambda \bar{W}_{i+1}\} / \mu_i , \\ \bar{N}_i &= \{(1-\delta_i)\lambda + (1-\delta_i)\lambda \bar{N}_{i+1}\} / \mu_i , \\ \bar{M}_i &= \{\delta_i \lambda + (1-\delta_i)\lambda \bar{M}_{i+1}\} / \mu_i . \end{aligned} \qquad (10)$$

(+) Subsequently, upper bar denotes mathematical expectation.

Henceforth, from (5) and (10) :

$$\bar{C}_i = \frac{(1-\delta_i)\lambda}{\mu_i} (\bar{C}_{i+1} + \gamma_n - \gamma_m) + \bar{C}^1_i, \qquad (11)$$

where $\bar{C}^1_i$ is the cost incurred on the average during period i when $\delta_i = 1$:

$$\bar{C}^1_i = (\gamma_d + i\gamma_w + \lambda\gamma_m) / \mu_i. \qquad (12)$$

Observe that the $\bar{C}^1_i$'s are "constant", that is, independent of δ; they are also readily computable from parameters of the problem.

Equation (11) holds for any diversion policy, hence also for an optimal one; denoting such a policy by $\tilde{\delta}$, and the corresponding costs by $\tilde{C}_i$, we then have :

$$\tilde{C}_i = \frac{(1-\tilde{\delta}_i)\lambda}{\mu_i} (\tilde{C}_{i+1} + \gamma_n - \gamma_m) + \bar{C}^1_i. \qquad (\tilde{11})$$

From this last equation, it is obvious that :

- if $(\tilde{C}_{i+1} + \gamma_n - \gamma_m) > 0$, then we must have $\tilde{\delta}_i = 1$, while
- if $(\tilde{C}_{i+1} + \gamma_n - \gamma_m) < 0$, then $\tilde{\delta}_i = 0$.

(This can also be justified by the principle of optimality: a car that arrives while i other ones are present causes a marginal cost of γ_m if it is diverted, and a cost $\tilde{C}_{i+1} + \gamma_n$ if it is not diverted).

Ignoring the cases where $(\tilde{C}_{i+1} + \gamma_n - \gamma_m) = 0$, (in these cases any value of δ_i is optimal), we may say that optimal diversion policies satisfy a property of all-or-nothing one encounters in other diversion assignment problems [01].

From $(\tilde{11})$, it is obvious that, since the $\tilde{C}_i$ are related to an optimal policy while the $\bar{C}^1_i$ concern a non-necessarily optimal one :

$$\tilde{C}_i \leqslant \bar{C}^1_i \quad , \quad \forall\ i.$$

Hence : $(\bar{C}^1_{i+1} + \gamma_n - \gamma_m) < 0$ implies, *a fortiori* : $(\tilde{C}_{i+1} + \gamma_n - \gamma_m) < 0$.

This leads to the following practical rule :

Proposition 1.

| If $\bar{C}'_{i+1} < \gamma_m - \gamma_n$, then $\tilde{\delta}_i = 0$.

Another case where a trivial solution to our optimization problem exists is that $\gamma_m - \gamma_n \leqslant 0$. Since $\tilde{C}_{i+1} > 0$, we then have :

Proposition 2.

| If $\gamma_m - \gamma_n \leqslant 0$, then $\tilde{\delta}_i = 1$, $\forall$ i.

In the remaining part of this section, we shall examine the case :

$$\gamma_m - \gamma_n > 0. \tag{13}$$

Proposition 3.

| $\tilde{C}_i \geqslant (\gamma_d + i\gamma_w + \lambda\gamma_n) / \mu_i$, $\forall$ i.

Proof.

$$\tilde{C}_i = \min_{\delta} \bar{C}_i \quad \text{(this minimum is taken under the constraints (2, 11, 12)}$$

$$\geqslant \min_{\delta_i} \frac{(1-\delta_i)\lambda}{\mu_i} \tilde{C}_{i+1} + \min_{\delta_i} \frac{(1-\delta_i)\lambda}{\mu_i} (\gamma_n - \gamma_m) + \bar{C}'_i \text{ , by (11)}$$

$$= 0 - \frac{\lambda}{\mu_i} (\gamma_m - \gamma_n) + \bar{C}'_i \text{ , by (13)}$$

$$= (\gamma_d + i\gamma_w + \lambda\gamma_n) / \mu_i, \text{ by (12)} \blacksquare$$

Propositions 1 and 3 are now used to establish

Theorem 1.

| Assume that the sequence $\{(\gamma_d + i\gamma_w + \lambda\gamma_n) / \mu_i ;\ i \geqslant n_2\}$ increases monotonously to infinity with i (hence the sequence $\{\bar{C}'_i\}$ also does).
| Let :

$$i_d = \max \{i \mid \bar{C}'_{i+1} < \gamma_m - \gamma_n\}. \tag{14}$$

| Then there exists an optimal diversion policy $\tilde{\delta}$ that satisfies :

$$\tilde{\delta}_i = \begin{cases} 0, \text{ for } i \leqslant i_d, & (15) \\ 1, \text{ for } i > i_d. & (16) \end{cases}$$

The basic assumption of this theorem can be justified as follows : at any time, the more there are trapped cars, the more they obstruct each other and the lower is the instantanous flow rate ($\mu_i \to 0$); in any way, this auto-obstruction implies that jam dissipation rates cannot be higher than that of non-congested M/M/∞ queues ($\mu_i = i\mu$); therefore μ_i never increases more rapidly than i ($\mu_i / i \to 0$).

Proof.

Equation (15) is a mere restatement of Proposition 1.

For proving (16), we proceed by contradiction and assume that :

$$\tilde{\delta}_j \neq 1, \text{ for some } j > i_d, \tag{17}$$

for any optimal policy $\tilde{\delta}$. Then we show that :

$$\tilde{\delta}_k \neq 1, \ \forall\ k \geqslant j. \tag{18}$$

But, using ($\tilde{11}$) with i = k, inequations (18) imply :

$$\tilde{C}_{k+1} + \gamma_n - \gamma_m \leqslant 0, \ \forall\ k \geqslant j \tag{19}$$

(otherwise, $\tilde{C}_{k+1} + \gamma_n - \gamma_m > 0$, and $\tilde{\delta}_k$ would be 1). This contradicts our basic assumption and Proposition 3, since $C_i \to \infty$ as $i \to \infty$.

It remains to establish (18), by induction.

If for some $k \geqslant i_d$: $\tilde{\delta}_k \neq 1$ for any optimal policy $\tilde{\delta}$, then :

$$\tilde{C}_{k+1} + \gamma_n - \gamma_m \leqslant 0.$$

But the case $\tilde{C}_{k+1} + \gamma_n - \gamma_m = 0$ is to be discarded, since $\tilde{\delta}_k$ may be chosen equal to 1. Therefore :

$$\tilde{C}_{k+1} + \gamma_n - \gamma_m < 0. \tag{20}$$

On the other hand, by definition (14) of i_d :

$$\bar{C}^1_{k+1} \geqslant \gamma_m - \gamma_n. \tag{21}$$

Inequalities (20) and (21) imply :

$$\bar{C}^1_{k+1} > \tilde{C}_{k+1}.$$

Reusing ($\tilde{11}$) with i = k+1 : $\tilde{\delta}_{k+1} \neq 1$ (otherwise, $\tilde{\delta}_{k+1} = 1$, and we would have $\tilde{C}_{k+1} = \bar{C}^1_{k+1}$) ■

CONCLUDING REMARKS

Theorem 1 provides a simple and practical rule for vehicular diversion since one can readily obtain the optimal *diversion level* from parameters of the congestion, and then use the policy defined by (15, 16). May be these parameters, and especially the jam dissipation rates, are not easy to estimate in practice, and this difficulty may limit the useness of the theorem. However, the kind of policy it states, — divert above some jam size —, seems very reasonable under natural conditions of auto-obstruction. Intuitively, such a result should hold also under more general hypotheses about arrival and departure processes. Note that the imbedded decision process owns the following trivial characteristic : choices must be made at and only at car arrival epochs. Therefore, even if the arrival process is "general" (that is, a renewal process) but the departure process is kept markovian, then the diversion decisional process remains markovian too, meaning that jam size at arrival times still constitutes the exhaustive information to our control problem.

Under the above stochastic assumptions, a recursive formula for $\bar{C}_i$, in an integral form, is not difficult to obtain. Unfortunately, we are not able to derive from it a generalization of Theorem 1.

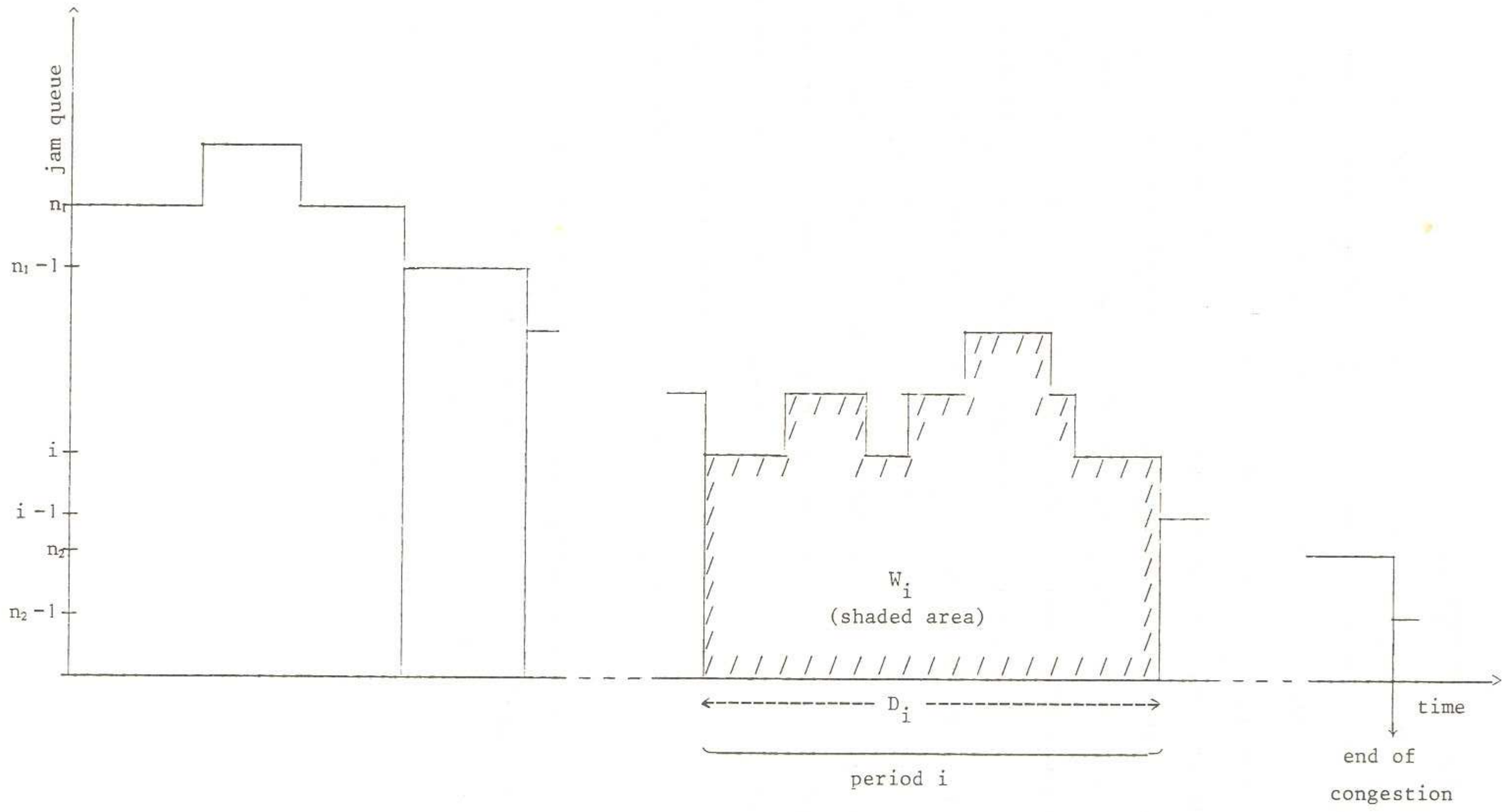


Figure 2. Decomposition of jam dissipation.

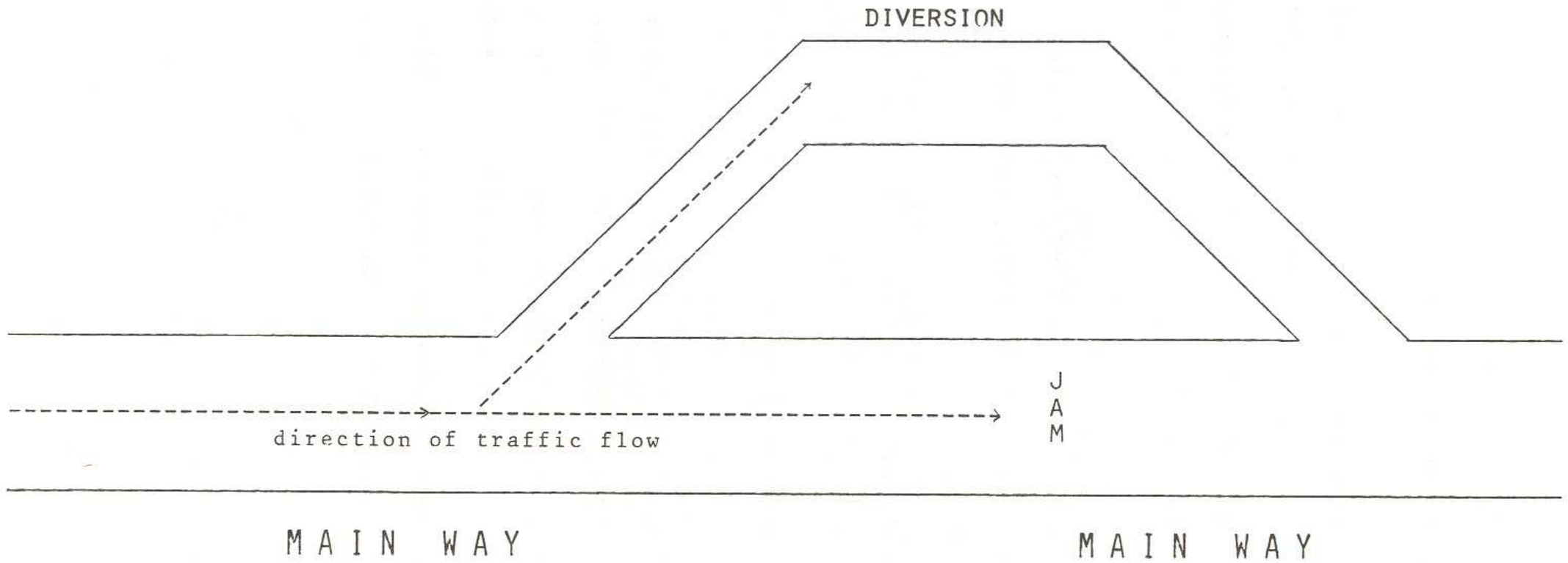


Figure 1. Map of the congested area.

REFERENCES

(Reference [08] is not cited in the text)

[01] Helly, W.,"Traffic Generation, Distribution, and Assignment," in Traffic Science, D. Gazis ed., Wiley & Sons, pp. 241-288 (1974)

[02] Agin, N.I. and Cullen, D.E., "An Algorithm for Transportation Routing and Vehicle Loading," in Logistics (vol. 1), M. Geisler ed., North-Holland, pp. 1-20 (1975).

[03] Duchateau, Ch. et Toint, Ph., "Un Modèle de Circulation pour 1'Agglomération Namuroise," Rep. 75/8:1, Fac. Univ. N.-D. de la Paix de Namur (1975).

[04] Darroch, J.N., Newell, G.F., and Morris, R.W.J., "Queues for a Vehicle-Actuated Traffic Light," Operation Res., pp.882-895(19

[05] Miller, A.J., "Settings for Fixed-Cycle Traffic Signals," Proc. 2nd Conf. Aust. Road Res. Board, vol.1, pp. 342-365(1964).

[06] McNeil, R. and Weiss, G.H., "Delay Problems for Isolated Intersections," in Traffic Science, D. Gazis ed., pp.109-174(197

[07] Gaver, D.P., "Highway Delays Resulting from Flow-Stopping Incidents," J. Appl. Prob., vol. 6, pp. 137-153 (1969).

[08] Natvig, B., "On the Waiting-Time and Busy Period Distributions for a General Birth-and-Death Queueing Model," J.Appl.Prob.,vol. pp. 524-532 (1975).

Belgian Journal of Operations Research, Statistics and Computer Science, Vol 25, n° 2-3.

TUTORIAL PAPER XIX

OPTIMAL CONTROL OF QUEUES: REMOVABLE SERVERS

J. TEGHEM Jr.

Département Gestion et Informatique
Faculté Polytechnique de Mons (*)
9, rue De Houdain, 7000 Mons
Belgium

ABSTRACT

This paper is devoted to the determination of the optimal operating rule for the behaviour of a removable server. We first examine the case of an individual service process, in M/G/1 queues with N-policy, D-policy and T-policy respectively; certain special cases are also examined such as the finite capacity case, the case of heterogeneous customers,... The problems with batch service are then described and two related applications are considered in details: the control of a shuttle and of a clearing system.

(*) This paper was written during a stay of visiting professor at Kyoto University (Department of Applied Mathematics, Faculty of Engineering) with the help of a grant from the Matsumae International Foundation.

Queueing systems have already a long life (see "Sixty years in queueing theory", Bhat, Mn. Sc. Vol.15, 6, pp.280) and after the second world war, queueing theory became not only a basic branch of applied probability theory, but also one of the classical methods of O.R. From the outset, some practical problems were treated by queueing systems : e.g., telephone exchange, job-scheduling, etc, ...; nevertheless, queueing theory is often considered by some operations research workers as "a fine mathematical model but ... inapplicable". One of the main reasons of such an opinion certainly is that queueing theory has examined interesting and sophisticated models in the field of applied probability, but often disconnected with decision or optimization problems of O.R. Although some control problems were early introduced in queueing models, they always were static or design problems (in which the system characteristics do not change over time) and clearly this type of problems did not meet the necessities of the greatest part of the practical queueing problems.

In the last twenty years, this situation has considerably changed in particular under the initial impulse, among others, of Naor'school in Israël and research workers, like Heyman, from Bell Telephone Loboratories in U.S.A. There has been an increasing interest in the study of dynamic control problems (in which the system characteristicsare allowed to change over time) and a lot of number of papers concerning this field have been published in the most famous journals of O.R.. One of the main reasons of this development surely is the existence of new practical queueing problems related to the management of great centers, in sectors like distribution, administration, public services, but especially to the management of computer centres. Moreover for the future, the present development of promising research fields, like queueing networks and the performance evaluation of computer networks, create new large possibilities for further applications of control models for more complex queueing systems. (For a stimulating review of such possibilities, see the report "Optimal control of admission to a queueing system" presented by Stidham Jr. at IFORS Congress (augustus 84)).

Several papers have already been published to review this new branch of queueing theory : Crabill-Gross-Magazine /73,77/ wrote a basic survey and a classified bibliography; Sobel /74/, Stidham Jr.- Prabhu /74/, Teghem Jr. /82/ presented others surveys; recently, some sections of Heyman-Sobel's book /82,84/, were devoted to this subject; a forthcoming invited review must soon be published in the European journal of O.R. (Teghem Jr. /85/).

According to the classification introduced by Crabill et al., we can distinguish four main categories of dynamic control models for a queueing system :

I. Control of the number of servers. The servers are removable : they may been turned on or off in function of the state of the system; the varying number of active servers must be determined.

II. Control of the service rate. This category often generalizes the first : the difference is rather than modifying the number of servers, the service process can be changed by varying the service rate.

III. Control of the admission of customers. In these problems, either the arrival rate can be modified, either customers can be refused; in some models the customers control themself the decision to enter into the system.

IV. Control of the queue discipline. Various papers deal with situations where the order of service can be determined. Generally, these problems concern either different classes of customers, either the allocation of customers to different servers.

In this paper, we only treat the first category; it is expected that we present soon, in this journal, other tutorial papers to cover the whole set of these models.

INTRODUCTION

The two first queueing models with a variable number of servers appearing in the literature are those of Romani /57/ and Moder-Philipps Jr. /62/; nevertheless, these papers are descriptive and no cost functions are introduced, a fortiori no optimization problems are setted. So we can consider that the real study of this type of problems begins with the paper of Yadin-Naor /63/. The classical *cost structure* related to a removable server consists in three types of cost :

- r_1 a non negative cost per unit time when the server is on, i.e. in activity; $r_2(\geq r_1)$ a non negative cost per unit time when the server is off, i.e. has decided to not be in activity; let us note $r=r_2-r_1$
- $R_1(R_2)$ a non negative fixed set-up (shut down) cost, incurred each time the server is turning on (off); let us note $R=R_1+R_2$ the total switching cost
- h a holding cost, or customer waiting cost, per unit time and per customer present in the system.

The *problem* consists to determine the optimal operating policies for the removable servers, i.e. to decide when open or close the service channels. It is clear that

- too long keeping a server on involves a too high running cost
- too long keeping a server off involves a too high holding cost
- too much times changing the state of the servers involves too high switching cost.

Thus the optimal policy must correspond to an equilibrium between these three situations.

The *review points*, i.e. the points at which the state of the server can be changed, are

(i) the arrival epochs : the server may be turned on

(ii) the service completion epochs : the server may be turned off.

Note : the hypothesis (ii) can be restrictive : see Heyman /68/, p.369.

The problem is a semi markovian decision process (SMDP) and the two classical criteria for SMDP, with infinite horizon time, can be introduced : either the discounted total cost, with a continuous discount factor $\beta<1$, either

the average cost per unit time. The theory of SMDP can be used to prove the existence of a non randomized stationary policy for this problem (see the books "Markovian decision processes" of Mine-Osaki (Elsevier 1970), "Dynamic programming" of Denardo (Prentice Hall 1982), but a special attention must be pointed out to the problem of unbounded costs (see Bell /71/, Lippman /73/, Stidham-Prabhu /74/).

I. A SINGLE REMOVABLE SERVER

Generally, the authors consider the problem in an M/G/1/L queue :

. L the maximum number of customers present in the system; in the most part of the studies L will be infinite
. customers arrive according a Poisson process, λ the mean arrival rate
. the service times are independant identically distributed random variable with distribution function B(.), finite expectation E(S) and finite variance. We denote $\tilde{B}(.)$ the LST of B(.) and $\rho=\lambda.E(S)$; we suppose $\rho<1$ if $L=\infty$.

I.A. N-Policy

The most part of the studies characterize the state of the system by the number of customers present and the server applies a N-policy : the state of the server, on or off, will be fixed in function of the number of customers. As preliminaries, let us introduce a particular subset of policies, playing a major part in the following.

Definitions

. A (ν,N) *policy*, with $0\leq\nu<N<L+1$, consists to turn the server on when N customers are present and turn it off when a service terminates with ν customers left in the system.
. The *policy* (0,L+1) (or$(0,\infty)$ if $L=\infty$) consists to always close the station.
. The *policy* (0,0) consists to always open the station.

Let us suppose that the server applies a (ν,N) policy. The different steady states of the system are

- $(i,0)$, with $\nu\leq i<N$: the server is off and there are i customers in the queue
- $(i,1)$, with $\nu+1\leq i<L+1$: the server is on and there are i customers in the system.

We will note

- $p_{ik}^{L}(\nu,N)$ the steady-state probability that the system is in state (i,k)
- $p_{i}^{L}(\nu,N)$ the steady-state probability that there are i customers in the system
- $p_{.0}^{L}(\nu,N) = \sum_{i=\nu}^{N-1} p_{i0}^{L}(\nu,N)$ the stationary probability that the server is off
- $N_{s}^{L}(\nu,N)$ the mean number of customers present in the system
- $\alpha^{L}(\nu,N)$ the mean busy period, i.e. a time interval beginning when the station is set up and terminating when for the first time thereafter, the number of customers in the system is equal to ν.
- $n_{b}^{L}(\nu,N)$ the mean number per unit time of busy cycles, composed of a successive busy period and idle period (i.e. a time interval during which the server is off).

Note We shall omit the indice L when $L=\infty$

Yadin-Naor /63/ have the first introduced the (0,N) policies in an infinite capacity, proving in particular that

$$N_s(0,N) = N_s(0,0) + \frac{N-1}{2} \tag{1}$$

$$p_{.0}(0,N) = 1-\rho \tag{2}$$

and Teghem Jr. /76/ has established a further relation between policies (0,N) and (0,0) :

$$\begin{aligned} p_i(0,N) &= \frac{1}{N} \sum_{j=0}^{i} p_j(0,0) && i<N \\ p_i(0,N) &= \frac{1}{N} \sum_{j=0}^{N-1} p_{i-j}(0,0) && i\geq N \end{aligned} \tag{3}$$

Loris-Teghem /82/ considered the case of a (ν,N) policy for a finite capacity and obtained, using some results of the theory of regenerative process, that

$$p^L_{i0}(\nu,N) = \frac{1}{N-\nu+\lambda\alpha^L(\nu,N)} \qquad \nu \leq i \leq N-1$$

$$p^L_{i1}(\nu,N) = \frac{\alpha^{i+1-\nu}(0,0)-E(S)}{E(S)(N-\nu+\lambda\alpha^L(\nu,N))} \qquad \nu+1 \leq i \leq N-1$$

$$p^L_{i1}(\nu,N) = \frac{\alpha^{i+1}(\nu,N)-\alpha^{i}(\nu,N)}{E(S)(N-\nu+\lambda\alpha^L(\nu,N))} \qquad N \leq i < L \tag{4}$$

$$p^L_{L1}(\nu,N) = \frac{(N-\nu)E(S)-(1-\rho)\alpha^L(\nu,N)}{E(S)(N-\nu+\lambda\alpha^L(\nu,N))}$$

As the mean idle period is obviously equal to $\frac{N-\nu}{\lambda}$, we have by an elementary renewal argument

$$n^L_b(\nu,N) = \frac{1}{\frac{N-\nu}{\lambda} + \alpha^L(\nu,N)} = \frac{\lambda p_{.0}(\nu,N)}{N-\nu} \tag{5}$$

1. The average case for M/G/1/L

For this criterion, the optimal stationnary policy is independant of the starting state

α) L=∞

Heyman /68/ proves the next property.

Property 1 [The optimal policy is either a policy (0,N) with $1 \leq N < \infty$, either the policy (0,0)

If C(N) denotes the average cost when the server applies a (0,N) policy, we have for $1 \leq N < \infty$.

$$C(N) = r_1 . p_{.0}(0,N) + r_2(1-p_{.0}(0,N)) + R.n_b(0,N) + h.N_s(0,N)$$

and using (1), (2) and (5).

$$C(N) = r_1 + r_2(1-\rho) + \frac{R\lambda(1-\rho)}{N} + h(N_s(0,0) + \frac{N-1}{2}).$$

For policy (0,0) we have

$$C(0) = r_2 + h\, N_s(0,0)$$

As C(N) is a convex function of N, Heyman /68/ concludes that the optimal value of N is either N=0, either one of the two integers surrounding the value

$$N^{*} = \sqrt{\frac{2\lambda R(1-\rho)}{h}} \qquad (6)$$

Remarks

(i) Three papers investigated this problem in the more general GI/G/1 queue. Sobel /69/ obtains sufficient conditions on more general cost structure for the existence of an optimal (ν,N) policy; Heyman-Marshall /68/ give bounds on the cost function and the optimal policy in the case of interrarrival distribution with increasing rate. Applying a method of diffusion approximation, Kimura-Ohno-Mine /80/ characterize the average cost rate and give some sufficient conditions under which the optimal operating policy falls into specific forms.

(ii) Talman /79/ gives another proof of the optimality of a (0,N) policy.

(iii) Yadin-Naor /63/ have also introduced the notion of set-up time, i.e. a random interval ellapsed before the service station is really reactived when such a decision is taken. For an M/M/1 queue, Baker /73/ analyses the consequences of an exponential set-up time on the value N^{*}.

β) $L<\infty$

Hersh-Brosh /80/ and Teghem Jr. /84/ have investigated the case $L<\infty$; in their model with limited capacity, the holding cost h is replaced by a penalty cost c_L incurred for every lost customer. When $L<\infty$, it is not more true that the policy (0, L+1) is never optimal. Using (4) and (5), Teghem Jr. /84/ extends and generalizes the results of Hersh-Brosh /80/ proving the next property :

Property 2

Consider the plane ($\overline{r} = \frac{r}{c_L}$, $\overline{R} = \frac{\lambda R}{c_L}$); it is divided in (L+2) regions corresponding to the optimality of policy (0,N), N=0,...,L+1, as showed in figure 1

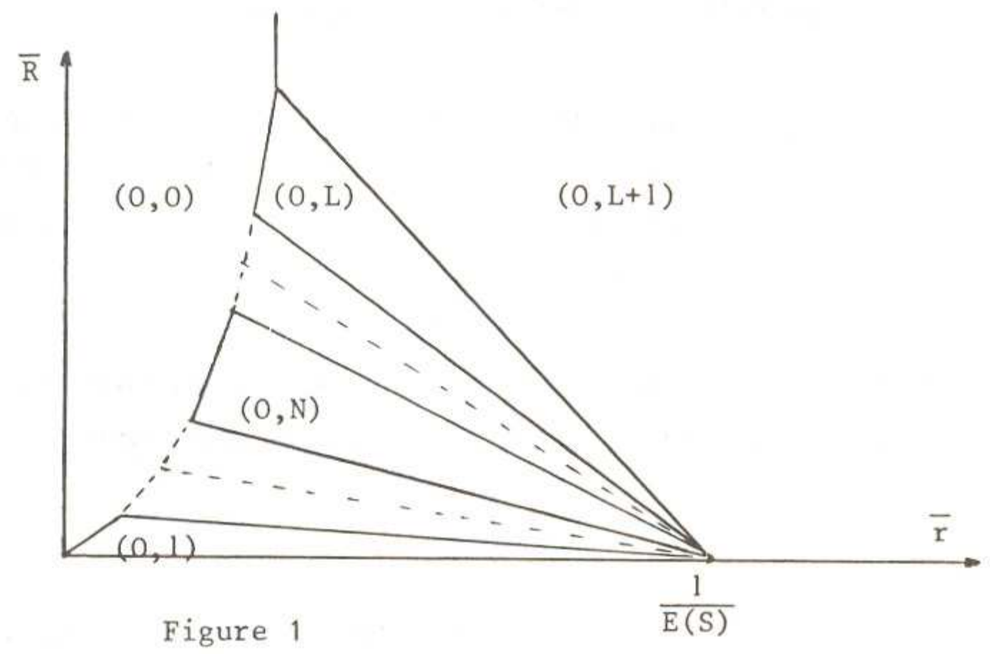


Figure 1

Moreover the equations of the frontiers of each region are explicited; they are also determined by the points of coordinates (r_N, R_N) with

$$r_n = \frac{1}{E(S)} \left(1 - \frac{1 + \lambda\, \alpha^{L-N}(0,0)}{1 + \lambda \alpha^{L}(0,0)}\right)$$

$$R_N = \sum_{j=0}^{N-1} (r_N - r_j)$$

Remarks

(i) The interaction between the optimal operating rule of a removable server in a finite capacity M/M/1 queue and the optimal behaviour of customers are simultaneously analysed in Teghem Jr. /77/

(ii) Bidhi Singh /82/ study a M/M/1/L queue, wherein, when the queue length increases to N(0<N<L), a search for an additional service facility for the service of a group of units is started; the availability time of this additional service facility is a random variable, but the search is dropped when the queue length reduces to some tolerable size ν. The optimal (ν,N) policy is investigated for a cost structure (with R=0) for this additional removable server.

2. The discounted case for M/G/1

For this criterion, the optimal stationary operating policy depends on the initial starting state; for easyness, let us suppose that this starting state is (0,0).

Heyman /68/ and Bell /71/ obtain the property

Property 3 — The optimal policy is either a policy (0,N), with $0 \leq N \leq \infty$, either a policy $(\overline{0,N})$, with $1 \leq N < \infty$, consisting to turn the server on at the first time when N customers are present and never off again.

Let us note $C_\beta(N)$ and $\overline{C}_\beta(N)$ the total discounted case, respectively for policies (0,N) and $(\overline{0,N})$. It is easy to determine

$$C_\beta(\infty) = \frac{r_1}{\beta} + \frac{h\lambda}{\beta^2}$$

but the determination of $C_\beta(N)$ and $\overline{C}_\beta(N)$ is more difficult. Heyman /68/ and Bell /71/ obtain

$$C_\beta(N) = (R_1 \tilde{A}_N + \frac{r_1}{\beta}(1-\tilde{A}_N) + \frac{r_2}{\beta} \tilde{A}_N(1-\tilde{G}_N) + R_2 \tilde{A}_N \tilde{G}_N + H(N))(1-\tilde{A}_N \tilde{G}_N)^{-1}$$

$$\overline{C}_\beta(N) = R_1 \tilde{A}_N + \frac{r_1}{\beta}(1-\tilde{A}_N) + \frac{r_2}{\beta} \tilde{A}_N(1-\tilde{G}_N) + \overline{H}(N)$$

where

. $\tilde{A}_N$ is the LST of the distribution of an idle period

$$\tilde{A}_N = (\frac{\lambda}{\lambda+\beta})^N \text{ for } N \geq 1 \quad \text{and} \quad \tilde{A}_0 = \frac{\lambda}{\lambda+\beta}$$

. $\tilde{G}_N$ is the LST of the distribution of a busy period, determined by

$$\tilde{G}_N = \tilde{G}(\beta)^N \text{ for } N \geq 1 \quad \text{and} \quad \tilde{G}_0 = \tilde{G}(\beta)$$

with $\tilde{G}(\beta)$ determined by the implicit equation

$$\tilde{G}(\beta) = \tilde{B}(\beta+\lambda-\lambda\tilde{G}(\beta)) \qquad (7)$$

. H(N) and $\overline{H}(N)$ are the terms corresponding to the holding costs.

Unfortunately, the expression of H(N) and $\overline{H}(N)$ are more complicated (see Bell /71/ p.210 and appendix) and, like $\tilde{G}(\beta)$,given by (7), almost impossible to calculate explicitely. Thus in practice, it is quite difficult to determine the optimal operating policy by the algorithm provided by Bell /71/. In order to deal with this difficulty, Kimura /81/ considers a diffusion approximation model depending only on the first two moments of the distribution function B(.)and derives approximation formula for $C_\beta(N)$ and $\overline{C}_\beta(N)$.

Remarks

(i) Blackburn /72/considers the same model, but with balking and two different types of reneging (single and batch reneging); for this case of impatient customers, he generalizes the results of property 3.

(ii) Langen /76/ gives a different form of the costs $C_\beta(N)$ and $\overline{C}_\beta(N)$.

3. Finite source M/G/1

α) Jaiswal-Simha /72/ consider a server applying a(0,N) policy in a finite source queue M/G/1 (i.e. each unit stays in the source for a random time exponentially distributed with parameter λ); let us note I the size of the source. In this model, which can be interpreted as a repair shop for I machines, the holding cost h is replaced by a reward g per unit time for each running machine, thus for each customer not present in the queueing system. These authors treat discounted, as well as undiscounted criterion; we give here, for instance, the average case.

Let P(N) be the total expected profit per unit time when the server applies a (0,N) policy; with an evident extension of notation, it is described by

$$P(N) = g(I-\overline{N}_s(0,N)) - R.\overline{n}_b(0,N) - r_1.\overline{p}_{.0}(0,N) - r_2(1-\overline{p}_{.0}(0,N))$$

when, for this finite source model, these authors obtain

$$\overline{N}_s(0,N) = I - \frac{1}{\lambda E(S)}(1-\overline{p}_{.0}(0,N))$$

$$\overline{n}_b(0,N) = \frac{\overline{p}_{.0}(0,N)}{\sum_{j=0}^{N-1} \frac{1}{(I-j)\lambda}}$$

with a specific value of $\overline{p}_{.0}(0,N)$ (see formula 22, p.701, Jaiswal-Simha /72/). Some numerical examples are treated for the determination of the optimal value of N.

β) We will introduce in this section a quiet different but interesting model concerning a simple closed queueing network. Hatoyama /77/ considers a *discrete time maintenance system* with I machines and two stations : an operating and a repair facility. The model is illustrated in figure 2 :

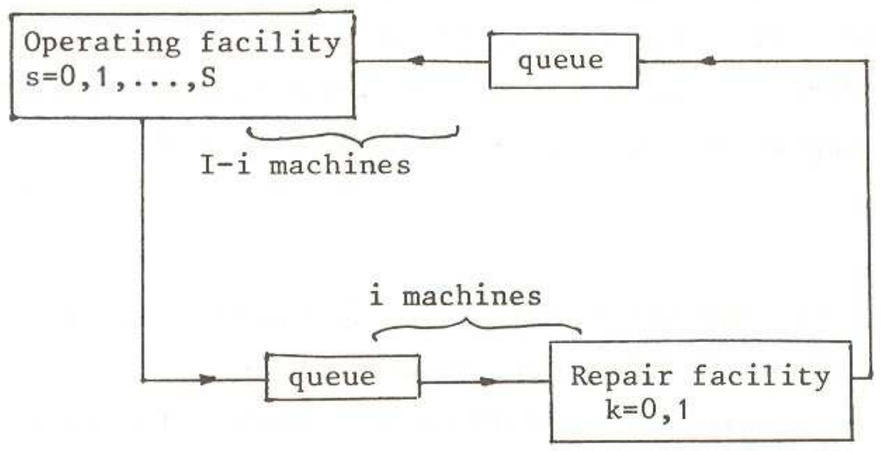


Figure 2

. Operating station : at the beginning of each period, an operating machine is classified as being in one of S+1 states (s=0,...,S), showing the degree of deterioration (0 : best state; S : failed state). An operating machine evolves from state s to state s' in one period , according to a transition probability $p_{ss'}$. An operating machine can be sent to the repair shop at any period and is then instantly replaced by a spare unit, if any available.

. Repair station : a machine sent to the repair shop must wait until all the machines which have already arrived at the repair shop are completely repaired; moreover, at the beginning of each period, the decision maker has the option of opening or closing the repair shop. When the repair station is open and there are i machines, q_{ij} is the probability that j of these machines are still in the repair system at the end of the period.
At the beginning of a period, the state of the system is thus described by (i,k; s) with
- i : the number of machines at the repair shop (i=0,...,I)
- k : equal to zero (one) if the repair shop is closed (open)
- s : the state of the machine at the operating station (when i<I).

This author associates with this system the following costs :

. Repair station : fixed switching costs R_1 and R_2; running cost r_1 per period; a general holding cost H(i,k) per period, depending of the state of the station.

. Operating station : an operating cost a(s) per period for a machine in state s; a reparing fixed cost c(s) for a machine in state s; a penalty cost P per period when no operating machine is available.

Hatoyama /77/ derives sufficient conditions on this structure cost for the existence of an optimal *two dimensional control-limit* policy : a control limit policy

- with respect to operating station : $\forall(i,k), \exists\, S_{(i,k)} \succ \forall s<S_{(i,k)}$ it is optimal to leave the operating machine and otherwise to repair it
- with respect to repair station : $\forall(s,k), \exists\, I_{(s,k)} \succ \forall i<I_{(s,k)}$ it is optimal to close the repair shop and otherwise to open it.

4. Several classes of customers

α) For the average case, Bell /73/ studies the optimal behaviour of a removable server in an M/G/1 priority queue, with two types of customers (k=1,2) having identical service time distribution, but characterized by different arrival rates λ_k and holding costs h_k. Customers of class 1 have non preemptive priority over customers of class 2 and $h_1 \geq h_2$.
Bell /73/ proves the property 4.

Property 4 — The optimal policy is either a policy $(0,N(i_1, i_2))$ consisting of turning the server off when the system is empty and turning the server on when i_1 or i_2, the number of customers of each class, reaches or crosses a linear boundary of the form $c_1 i_1 + c_2 i_2 + d = 0$; either the policy (0,0).

This author establishes that the line $c_1 i_1 + c_2 i_2 + d = 0$ is included, like in figure 3, between the two lines $i_1 + i_2 = N_k^{\ast\ast}$ (k=1,2) corresponding to the cases of a same holding cost h for each customer, respectively equal to h_1 and h_2; moreover its slope is always smaller than -1 and equal to -1 only when $h_1=h_2$.

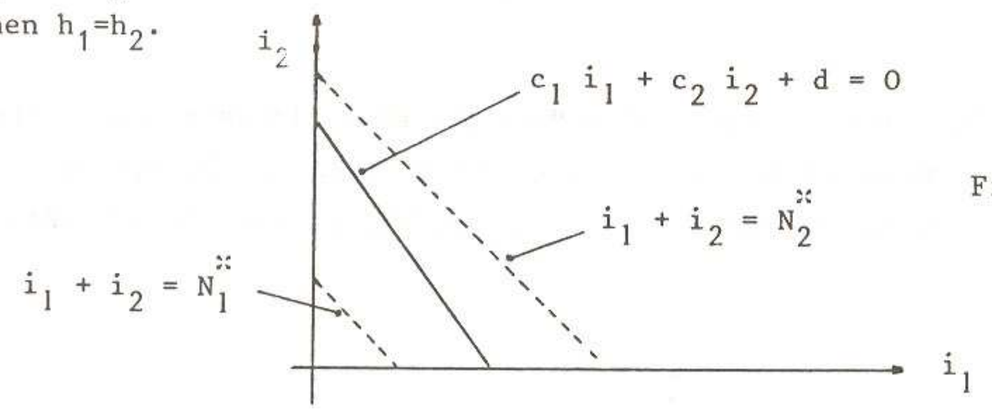


Figure 3

Let us note that Tijms /74/ extends these results to the case of different service time distributions for each type of customers and derives, using some results of the theory of regenerative process, expression for the average number of customers, of each class, present in the system.

β) In his Ph.D.Ghorayeb /78/ considers the same type of model but without any assumption concerning the priority and with switching costs, also to move the server from one class to the other. The operating policy must then determine not only when to open or to close the station, but also which class of customers to serve and when to change to the other class. This general problem seems really difficult and the author introduces some limitative assumptions : the main one is that there are no arrivals of class 2 when the server is busy. Ghorayeb /78/ proves property 5.

Property 5

There exists an optimal policy represented by figure 4 in the plane (i_1, i_2) and such that :

- . in area I : if the server is off or on, he remains off or on
- . in area II : if the server is off, he is turned on and he begins to serve first the customers of class 1
- . in area A : if the server is on, he continues to serve the same type of customer
- . in area B : if the server is on, he always serves customers of class 1.

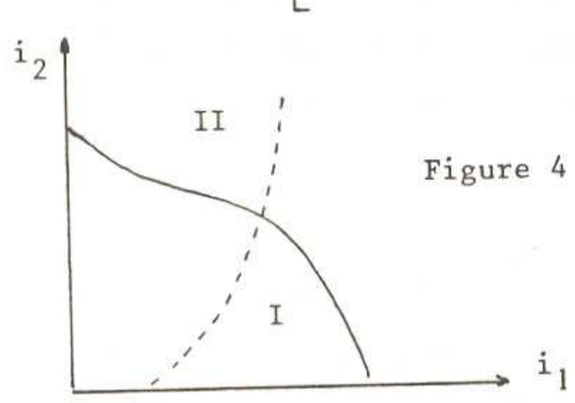
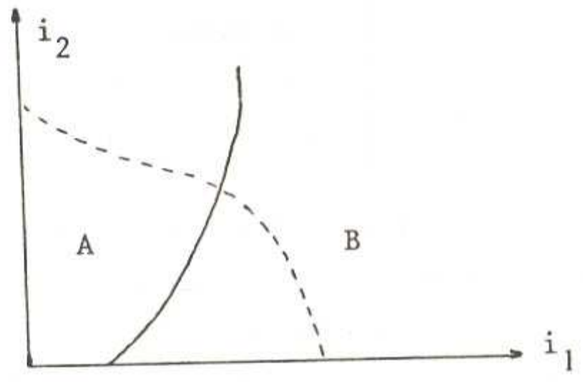


Figure 4



I.B. *D-Policy*

Balachandran /73/ has the first introduce the model in which the state of the system is the workload i.e. the total amount of work in the system. The idea is that the customer's service times are different even though they

may come from the same distribution; yet this type of measure means that service times must be known immediately, after the customer enters the queue. A(O,D) policy consists then to turn on the server when the total work to be done reaches the value D and turn him off when the system is empty. For the average cost criterion, this author analyses the (O,D) policies for a similar cost structure as in I.A., except that $r_1=r_2=0$ and h is now a holding cost per unit time per unit work. If C(D) represents the average cost, we have like in I.A., and with an evident extension of notations,

$$C(D) = R.n_b\ (O,D) + h\ W(O,D)$$

where W(O,D) is the expected work in the system.
Balachandran-Tijms /75/ and Tijms /76/ obtain

$$n_b(O,D) = \frac{\lambda(1-\rho)}{E(M_D)}$$

$$W(O,D) = W(O,O) + D - \int_0^D \frac{E(M_x)}{E(M_D)}\ dx$$

where M_x represents the number of customers present at the opening of the station if a (O,x) policy is applied and W(O,O) is the average waiting time in an M/G/1 system with policy (O,O); these authors derive D^* the value corresponding to the minimum of C(D)

$$D^* + \int_0^{D^*} E(M_x)\ dx = \frac{R\ \lambda(1-\rho)}{h}$$

For the cost C(D), the policy (O,D^*) is compared with the policy (O,N^*).
For a constant service time, the two policies are obviously equivalent.
Boxma /76/ generalizes results of Balachandran-Tijms and proves the optimality of the D policy over the N policy. Note that Tijms /77/ gives an expression of the stationary distribution of the workload, when a policy (O,D) or a policy (O,N) is applied.

I.C. *T-Policy*

1. FIFO discipline

Levy-Yechiali /75/ consider an M/G/1 queue, with usual FIFO discipline

("First in first out"), such that when the server finishes serving a unit and finds the system empty, he goes away for a length of time called a vacation. At the end of the vacation, the server returns to the main system and begins to serve if they are customers. If the server finds the system empty at the end of a vacation, two models are introduced :

Model 1 : the server waits for the first customer to arrive and then an ordinary busy period begins

Model 2 : the server immediately takes another vacation and continues in this manner until he finds at least one waiting unit upon return from a vacation.

As usual, the server serves the queue as long there is at least one unit in the system.

F(.) denotes the distribution function of the random vacation T, with finite mean E(T) and second moment $E(T^2)$. The same cost structure than in section I.A. is introduced, except that $r_2 < r_1$ and in fact $r=r_1-r_2>0$ represents the reward per unit time of the server for the work done during the vacation.

By evident extension of notation, the average profit per unit time for models j=1, 2 respectively, is equal to

$$P^{(j)}(T) = r\, p_{.0}^{(j)}(0,T) - R.n_b^{(j)}(0,T) - h.N_s^{(j)}(0,T)$$

Let we note

$$\gamma = \frac{\lambda E(T)}{f_0+\lambda E(T)} \qquad \text{with } f_0 = \int_0^\infty e^{-\lambda t} dF(t)$$

Levy-Yechiali /75/ obtain

$$p_{.0}^{(1)}(0,T) = \gamma . p_{.0}^{(2)}(0,T) \qquad \text{with } p_{.0}^{(2)}(0,T) = 1-\rho$$

$$n_b^{(1)}(0,T) = \frac{\gamma}{1-f_0} \cdot n_b^{(2)}(0,T) \qquad \text{with } n_b^{(2)}(0,T) = \frac{(1-f_0)(1-\rho)}{E(T)}$$

$$N_s^{(1)}(0,T) - N_s(0,0) = \gamma(N_s^{(2)}(0,T) - N_s(0,0))$$

with $$N_s^{(2)}(0,T) - N_s(0,0) = \frac{\lambda E(T^2)}{2E(T)}$$

so that $$P^{(1)}(0,T) = \gamma(P^{(2)}(0,T) - R\,\frac{(1-\rho)f_0}{E(T)})$$

These authors conclude that, for a fixed distribution F(.), model 2 is superior

to model 1 if and only if

$$P^{(2)}(0,T) > -\lambda(1-\rho)R;$$

they also determine the expressions of an optimal value of T (his expected value) for deterministic (exponential) vacation times.

Note As $p_0^{(2)}(0,T)$ is independant of T, there is no need to introduce the cost r in model 2.

Heyman /77/ examines model 2; when the server finds the system empty at the end of a vacation, he considers that a busy period of length zero occurs and that the cost R is thus incurred; in that case we have

$$n_b^{(2)}(0,T) = \frac{1-\rho}{E(T)}$$

and for deterministic vacation times, the optimal value of T is

$$T^* = \sqrt{\frac{2\ R(1-\rho)}{\lambda h}} = \frac{N^*}{\lambda}$$

This author compares the average cost for this optimal T policy and the optimal N policy and proves that the latter always does better than the former.

Remarks

(i) Meilijson-Yechiali /77/ consider a priority control model in a GI/G/1 queue, in which insertion of idle time is allowed.

(ii) Van der duyn Schouten /78/ introduces a descriptive model with stochastic vacation time and a finite capacity for the workload and derives several characteristics : the joint stationary distribution of the workload and the stage of the server; the average number of overflows per unit time and the average number of vacations per unit time.

(iii) Note that some queueing problems in which the service station is subject to breakdown are close of the removable server model.

2. SPT discipline

α) Three papers have been more recently published by Shanthikumar /80[a],80[b],81/; note that his results can be applied as well to N-policy that to T-policy, but

we only present the latter. In the first paper, the author develops a new and interesting method to analyze some controlled M/G/1 queueing problems, using properties of the number of up and downcrossings levels in a special case of regenerative process. He obtains two important basic relations between the density and the expected number of upcrossings of this regenerative process (see formula 8 and 9, p.817, Stanthikumar /80[a]/); these equations can be used in many queueing systems,especially with exponential arrivals. For instance, Stanthikumar /80[a]/ uses this method to easily derive the results of Levy-Yechiali /75/ concerning the virtual waiting time distribution for T-policy.

β) By this method, Shanthikumar /80[b]/ analyses optimal T-policy (model 2) of a server in an M/G/1 queue with shortest processing time (SPT) discipline : at the service completion epochs, the server choosesto serve the customer with the shortest service time. For this model, let us note W(SPT; T) the expected waiting time of an arbitrary customer; W(FIFO; T) may be determined by relation (8) and Little formula.
Shanthikumar /80[b]/ determines the LST of the waiting time distribution; then he obtains W(SPT; T) and proves the next conservation identity :

$$\frac{W(FIFO;T)}{W(SPT;T)} = \frac{W(FIFO;0)}{W(SPT;0)} \qquad \forall\ T \qquad (10)$$

In the case of deterministic vacation time, he derives the optimal value of T

$$T^{*}(SPT) = T^{*}(FIFO)\ .\ \sqrt{E} \qquad (11)$$

where $T^{*}(FIFO)$ is given by (9) and E is the value of identity (10).

γ) Shanthikumar /81/ applies the same procedure for a different, but quite close, queueing discipline, called SPT within generations.
(SPT-WG) : the customers that arrive during the vacation form the first generation and their total service time is the lifetime of the generation; customers arriving during the lifetime of the first generation, if any, make up the second generation, with its lifetime, and so on; within each generation, customers are served in the order of the SPT discipline.
This author obtains similar results as (10) and (11), i.e. with obvious extension of notation

$$\frac{W(FIFO;T)}{W(SPT\text{-}WG;T)} = \frac{W(FIFO;0)}{W(SPT\text{-}WG;0)} = E' \qquad \forall\ T$$

and

$$T^{*}(SPT\text{-}WG) = T^{*}(FIFO)\ .\ \sqrt{E'}$$

II. MULTI REMOVABLE SERVERS

The problem of more than one removable server is only investigated in a few studies. A basic difference is that when the unique server is turned off, the queue size necessarily must increase, but when one of several servers is turned off, the queue size may go up and down. Mc Gill /69/ was the first to examine this problem and established some intuitive properties for the form of optimal policies, in the case of a general discounted cost GI/G/1 queueing system, but only for a finite horizon. Bell /75/ considers this problem for an infinite horizon M/M/S model; a classical cost structure is considered and fixed switching costs are incurred to turn each server on or off. The state of the system is now denoted (i,k) when there are i customers and k servers working. This author calls *efficient policy*, an operating rule which never allows more working servers that customers present; otherwise the policy will be called inefficient. He proves that for r sufficiently high and all others parameters fixed, there exists an efficient policy; otherwise an optimal policy may turn one or more servers off, even when there are customers for him to serve, i.e. may be inefficient. Bell /80/ further investigates this model for S=2 and first shows that a critical number N exists such that all the servers should be turned on or left on in any state (i,k) with $i \geqslant N$. Generalizing (ν,N) policies of section I, he definies a (ν_1,ν_2,N_1,N_2) policy for which the 4 critical levels denote numbers of customers in the system when the number of working servers should be adjusted downward to 0,1 and upward to 1,2 respectively. For R=0, obviously an optimal policy adjusts the number of working servers to min {i,k}, i.e. $\nu_1=0$, $\nu_2=N_1=1$, $N_2=2$.
If R is allowed to increase from 0, he obtains the following property.

Property 6 — The best efficient policies is such that $\nu_1=0$, $1<\nu_2$.
Yet, if an inefficient policy is optimal, it may be of three types :

- leave both servers on at all times ($\nu_1=\nu_2=N_1=N_2=0$)
- leave at least one server at all times ($\nu_1=N_1=0$)
- decrease the number of servers only in state (0,2) and turn off both servers ($\nu_1=\nu_2=0$).

Remarks

(i) Magazine /69/ and Huang - Brumelle - Sawaki - Vertinsky /77/ consider control models under periodic reviews, i.e. the review points are at equally spaced time intervals.

(ii) Levy-Yechiali /76/ consider T-policies in an M/M/S queueing system and derives formula for the distribution of the number of busy servers and the mean number of units in system.

(iii) Winston /78/ examines several removable servers in an exponential queueing system in which the arrival rate depends on the number of customers. For state (i,k), a general holding cost h(i) and a running cost r(k) are introduced, but no switching costs. This author derives conditions that ensure the optimality of monotone policies such that the number of working servers is a non decreasing function of the number of customers in the system.

III. BATCH SERVICE AND RELATED AEREAS

III.A. Batch service

An interesting problem concerns a removable server who can make a decision to serve any number of customers in a batch, up to some batch size limit $1\leq Q\leq\infty$. For this model, Deb /76/ introduces the following cost structure :

- r_1, r_2, R_1, R_2 like before
- h(i) : a general non linear cost for holding i customers during a unit time
- c.y a linear cost for serving y customers (we have y=min(i,Q)).

The approach of this author is different in the sense that he establishes the form of an optimal policy by direct analysis of the infinite horizon

functional equation of SMDP.

Let us introduce relations (11) and (12) respectively for the discounted and average criterion :

for some $\eta>0$, $\quad h(i) - h(i-1) > \eta \qquad (12)$

$$h(i) - h(i-1) > \frac{1-\tilde{B}(\beta)}{\tilde{B}(\beta)Q}(\beta R_1 + r) + \frac{\beta}{\tilde{B}(\beta)} c + \eta \qquad (13)$$

Property 7

a) If (12) and (13) hold, respectively for the two criteria, there exists an optimal policy of the form (ν,N)

b) Otherwise the policy $(0,\infty)$ of turning the server off for ever is optimal.

Remark

In the particular case $R_1=R_2=0$, Deb-Serfozo /72/ proves that for property 7.a), we have $\nu=N-1$, i.e. there exists an optimal policy of the type control limit policy. For this case and moreover with $Q=\infty$, Weiss /79,81/ presents some properties of the cost-function, gives an algorithm for finding the optimal control limit and determines the waiting time distribution.
Finally, let us note that Weiss-Pliska /82/ introduce a general holding cost $h_t(i)$ depending of time and show that control limit policies may cease to become optimal.

III.B. Control of a shuttle

Batch service queueing systems are often found in transportation, since mass transit vehicles are natural batch servers; so a related application of model described in III.A. is the optimal dispatching of a shuttle. Let be a shuttle system consisting of a single carrier with capacity Q, transporting passengers between two terminals. At each terminal, passengers arrive according to independant Poisson processes (λ_1,λ_2) and all arriving passengers wait to be transported to other terminals where they exit the system; B(.) is here the distribution of the interterminal travel times,

independant of everything and in particular of the load carrier. The system is reviewed at those points in time when, either the carrier has just arrived at one of the terminals or when the carrier is waiting at one of the terminals and a new passenger arrives. The state of the system is denoted by (i_0,i_1,δ), where i_0,i_1 are the number of passengers at the two terminals and δ, equal to zero or one, indicates at which terminal, 0 or 1, is the carrier. At each review point, it is necessary to decide if the carrier is dispatched (with $\min(i_\delta,Q)$ customers) or not. The cost of carrying y passengers is $R + c.y$ and there is a linear holding cost h.

1. Control at both terminals

Some particular cases have been first considered. Ignall-Kolesar /72/ study the case Q=1 and moreover the dispatch decision is made without knowledge of the queue at other terminal; the paper of Barnett /73/ concerns deterministic travel times with some restrictions on the values λ_1 and λ_2. Barnett-Kleitman /78/ show that the result for the control at a single terminal (see below) is not directly generalized for control at both terminals. The general model is introduced by Deb /78/. In a first time, he considers the finite horizon period n and extends then the results for $n\to\infty$; for the discounted criterion, he proves the following property.

Property 8

a) If $h<\beta(c+\frac{R}{Q})$, then the policy of never dispatching the server is optimal

b) Otherwise, the optimal control policy is of the form : Dispatch the carrier iff $i_\delta \geq G_\delta(i_{1-\delta})$, where $G_\delta(.)$ is a monotone decreasing control function.

Unfortunately, the explicit determination of the function $G_\delta(.)$ seems an unsolved problem.

2. Control at one terminal

Ignall-Kolesar /74/ examine the particular problem of the control at

a single terminal - said zero - and for an infinite capacity shuttle with deterministic travel times. They prove the following property :

Property 9 — There exists an optimal control limit policy of the form : Dispatch the carrier iff $i_0 + i_1 \geq N$.

Weiss /81/ presents a method for computing the control limit N, compares this policy with the more traditional policy of scheduled periodic service and last, proves a conjecture of Ignall-Kolesar /74/ regarding the case when the dispatcher does not know the number of passengers at terminal 1 : there exists an optimal control policy concerning i_0 plus the expected number of passengers at terminal 1.

Remarks

(i) Osuna-Newell /72/ and Asgharzadeh-Newell /78/ consider a particular model of multiple vehicle system.

(ii) Teghem Jr. /82/ consider a double shuttle system, like a ropeway, transporting passengers simultaneously from one terminal to the other in the two opposite directions.

III.C. Clearing systems

Stochastic clearing systems are first analysed by Stidham Jr. /74/ and optimized by the same author in 77.
The cumulative input to such a system is described by a non decreasing stochastic process $\{Y(t),\ t\geq 0\}$, with $Y(0)=0$; output occurs intermittently in the form of clearing operations, which instantaneously remove all the quantity in the system. This author considers that clearing occurs whenever the cumulative input since the last clearing instant exceeds a critical level q. Let us introduce some definitions and notations.

. X_1, time until first clearing; X_n, time between $(n-1)^{th}$ and n^{th} clearings $(n>1)$

. $S_0=0$, $S_n = \sum_{j=1}^{n} X_j$, $n\geq 1$

. $R(t) = \max\{n \mid S_n \leqslant t\}$, the number of clearings in $[0,t]$

. $V(t) = Y(t) - Y(S_{R(t)})$, the net quantity in the system

. $T(y) = \inf\{t \mid Y(t)>y\}$, the first entrance time into the set (y,∞)

. $W(y) = E(T(y))$, the sojourn measure of the set $[0,y]$.

With

Assumption 1 $\{V(t),t\geqslant 0\}$ is a regenerative process with respect to the renewal sequence X_n, $n\geqslant 1$.

Stidham Jr. /74/ obtains the stationary distribution of $\{V(t),t\geqslant 0\}$: it is completely defined by knowledge of $W(y)$, $\forall y\leqslant q$ and is different that the stationary distribution uniform between 0 and q.

Stidham Jr. /77/ introduces the following costs :

. a positive cost R - independant of q - whenever a clearing takes place

. a general holding cost $h(x) \geqslant 0$, incurred while $V(t)=x$

and

Assumption 2 h is continuous and - h is unimodal with mode x_0 $\varepsilon(-\infty,\infty)$. Let we note $h'(x) = h(x_0+x)$.

The average cost C(q) is given by

$$C(q) = \frac{R + \int_0^q h'(x-x_0)\, dW(x)}{W(q)}$$

and this author proves.

Property 10 [Let $\tilde{q}$ be a solution to the equation $\int_0^q W(x)dh'(x-x_0) = R$ (14)
Then $\tilde{q}$ minimizes C(q) among all $q\geqslant 0$, such that $W(q)>0$
(If there is no solution to (14), then $\tilde{q}=\infty$).

Rather than "N-policy", Nishimura /79/ considers T-policy in this model. He first obtains an optimal clearing interval $\tilde{T}$ among the set of non negative random variables with finite mean (see theorem 3.3., p.101, Nishimura /79/) and then proves that if h(x) is continuous and monotone non decreasing in $x\geqslant 0$, then $\tilde{T} = T(\tilde{q})$.

These two authors Stidham Jr. /77/ and Nishimura /79/ generalize their results to a general clearing system in which the effect of a clearing operation is that the quantity in the system is restored to a level ν rather than 0.

Last, let us note that Whitt /81/ further investigates the comparison between the stationary distribution of V(t) and the uniform distribution and Stidham-Serfozo /73/ introduce more general clearing systems in which, in particular, the quantities cleared are random variables.

CONCLUSION

We have here examine some of the principal papers related to removable servers; yet we have of course no claim to be exhaustive. It is important to remark that there is no major difficulty to classify in categories the different papers because, unfortunately, very few studies consider the optimization of more than one parameter. It seems us important in the future to analyze the interactions between several different optimization problems. We invite the reader, interested by further comments on the prospects of the field of optimal control queueing problems, to refer to the forthcoming paper of Teghem Jr. /85/.

To conclude, we want to turn the attention of the reader on the possibility to determine an optimal policy of very complex optimal control problems - for which it can not be expected that the optimal policy has simple form - by using numerical algorithms issued of the SMDP theory. A good example of this technic is given by the paper of K.Ohno-K.Ichiki /84/ ("An optimal control problem of a C-stage tandem queueing system" Technical report-Department of Information processing and Management Sciences, Faculty of Science, Konan University, Kobe, Japan). It is important to not forget this type of approach to resolve complex practical problems.

CLASSIFIED BIBLIOGRAPHY

By easyness we shall refer to

(*) T.B.CRABILL, D.GROSS, M.J.MAGAZINE /77/ "A classified bibliography of research on optimal design and control of queues" ORSA, vol.25,2 (219-232)

SURVEY PAPERS

. T.B.CRABILL, D.GROSS, M.J.MAGAZINE /73/ see(*)

. D.P.HEYMAN, M.J.SOBEL /82,84/ "Stochastic models in Operations Research - volumes I and II" Mc Graw-Hill Book Company

. M.J.SOBEL /74/ see(*)

. S.STIDHAM Jr., N.U.PRABHU /74/ see(*)

. J.TEGHEM Jr. /82/ "New developments in Optimal control of queueing systems" in Operations Research in Progress (333-349), D.Reidel Publishing Company

. J.TEGHEM Jr. /85/ "Optimal control of queueing systems - Invited review", forthcoming in E.J.O.R.

I. A SINGLE REMOVABLE SERVER

I.A. N-Policy

. K.R.BAKER /73/ see(*)

. C.E.BELL /71/ see(*)

. C.E.BELL /73/ "Optimal operation of an M/G/1 priority queue with removable server" ORSA, vol.21, 6 (1281-1289)

. K.BIDHI SINGH /82/ "Optimal operating policy for a random additional service facility" J.O.R.S. vol.33 (837-843)

. J.D.BLACKBURN /72/ see(*)

. A.GHORAYEB /78/ "Politiques optimales dans un système d'attente avec 2 classes de clients et un guichet pouvant s'ouvrir et se fermer" Thèse de doctorat - Faculté des Sciences - Université Libre de Bruxelles

. Y.HATOYAMA /77/ "Markov maintenance models with control of queue" JORS Japan, vol.20,3 (p.164-181)

. M.HERSH, I.BROSH /80/ "The optimal strategy structure of an intermittently operated service channel" EJOR,5 (133-141)

. D.P.HEYMAN /68/ see(*)

. *D.P.HEYMAN, K.T.MARSHALL /68/* see(*)

. *N.K.JAISWAL, P.S.SIMHA* see(*)

. *T.KIMURA, K.OHNO, H.MINE /80/* "Approximate analysis of optimal operating policies for a GI/G/1 queueing system" Memoirs of the Faculty of Engineering, Kyoto University, vol.XLII, part 4

. *T.KIMURA /81/* "Optimal control of an M/G/1 queueing system with removable server via diffusion approximation" EJOR,8,4 (390-398)

. *H.J.LANGEN /76/* "Optimierung von M/G/1 - wartesystemen mit schaltbarem Bediener" - Proceedings Symposium of Heidelberg

. *S.A.LIPPMAN /73/* see(*)

. *J.LORIS-TEGHEM /84/* "Imbedded and non imbedded stationary distributions in a finite capacity queueing system with removable server" Cahiers du C.E.R.O. vol 26, n° 1-2, 1984 (87-94)

. *J.G.SHANTHIKUMAR /80a,80b,81/* see T.C.

. *M.J.SOBEL /69/* see(*)

. A.J.TALMAN 79/ "A simple proof of the optimality of the best N-policy in the M/G/1 queueing control problem with removable server " Statistics Neerlandica, 33-3 (143-150)

. *J.TEGHEM Jr. /76/* "Properties of (0,k) policy in a M/G/1 queue and optimal joining rules in a M/M/1 queue with removable server" in Operational Research'75 (229-259), North Holland Publishing Company

. *J.TEGHEM Jr. /77/* "Optimal pricing and operating policies in a queueing system" in Advances in Operations Research, (489-496), North Holland Publishing Company

. *J.TEGHEM Jr. /84/* "Optimal control of a removable server in an M/G/1 queue with finite capacity" Technical report-Department Gestion et Informatique - Faculté Polytechnique de Mons - Belgium

. *H.TIJMS /74/* see(*)

. *M.YADIN, P.NAOR /69/* see(*)

I.B. *D-Policy*

. *K.R.BALACHANDRAN /73/* see(*)

. *K.R.BALACHANDRAN, H. TIJMS /75/* see(*)

. O.J.BOXMA /76/ "On the D-policy for the M/G/1 queue" Mn.Sc. vol 21,9, (1073-1076)

. H.TIJMS /76/ "Optimal control of the workload in an M/G/1 queueing system with removable server" Math.Operations forsch.u. Statist.7 (933-943)

. H.TIJMS /77/ "Stationary distribution for control policies in an M/G/1 queue with removable server" Technical report - Vrije Universiteit Amsterdam

I.C. T-Policy

. Y.LEVY, U.YECHIALI /75/ "Utilization of idle time in an M/G/1 queueing system" Mn.Sc., vol 22, n°2 (202-211)

. D.P.HEYMAN /77/ "The T-policy for the M/G/1 queue" Mn.Sc. vol 23, n°7 (775-778)

. I.MEILIJSON, U.YECHIALI /77/ "On optimal right-of-way policies at a single server station when insertion of idle times is permitted" Stochastic processes and their applications, 6 (25-32)

. J.G.SHANTHIKUMAR /80[a]/ "Some analyses on the control of queues using level crossings of regenerative processes" J.Appl.Prob.17, (814-821)

. J.G.SHANTHIKUMAR /80[b]/ "Analysis of the control of queues with shortest processing time service discipline" JORS Japan, 23, 4 (341-352)

. J.G.SHANTHIKUMAR /81/ "M/G/1 queues with scheduling within generations and removable server" ORSA, 29, 5 (1010-1018)

. F.A. VAN DER DUYN SCHOUTEN /78/ "An M/G/1 queueing model with vacation times" Z.O.R. 22 (95-105)

II. MULTI REMOVABLE SERVERS

. C.E.BELL /75/ see(*)

. C.E.BELL /80/ "Optimal operation of an M/M/2 queue with removable servers" ORSA, 28, 5 (1189-1204)

. C.C.HUANG, S.L.BRUMELLE, K.SAWAKI, I.VERTINSKY /77/ "Optimal control of multi-servers queueing systems under periodic review" N.R.L.Q 24,1 (127-136)

. Y.LEVY, U.YECHIALI /76/ "An M/M/S queue with server's vacations" INFOR ,14,2 (153-163)

. M.J.MAGAZINE /71/ see(*)

. J.T.Mc GILL /69/ see(*)

. J.J.MODER,C.R.PHILIPPS Jr. /62/ see(*)

. J.ROMANI /57/ see(*)

. W.WINSTON /78/ "Optimality of monotonic policies for multiple server exponential queueing systems with state-dependant arrivals rates" ORSA, vol26,6 (1089-1094)

III. BATCH SERVICES AND RELATED AEREAS

III.A. Batch services

. R.K.DEB,R.F.SERFOZO /73/ see(*)

. R.K.DEB /76/ "Optimal control of batch service queues with switching costs" Adv.Appl.Prob.,8 (177-194)

. H.J.WEISS /79/ "The computation of optimal control limits for a queue with batch services" Mn.Sc. vol 25,4 (320-328)

. H.J.WEISS /81[a]/ "Waiting time distribution in a queue with a bulk service rule" Opsearch,18,1 (15-24)

. H.J.WEISS,S.R.PLISKA /82/ "Optimal policies for batch service queueing systems" Opsearch,19 (12-22)

III.B. Control of a shuttle

. K.ASGHARZADEH, G.F.NEWELL /78/ "Optimal dispatching strategies for vehicles having exponentially distributed trip times" NRLQ,25,3 (489-510)

. A.BARNETT /73/ "On operating a shuttle service" Networks 3 (305-313)

. A.BARNETT, D.J.KLEITMAN /78/ "On two terminal control of a shuttle service" SIAM J.Appl.,vol 35, 2 (229-234)

. R.K.DEB /78/ "Optimal dispatching of a finite capacity shuttle" Mn.Sc. vol 24, 13 (1362-1372)

. E.IGNALL, P.KOLESAR /72/ see(*)

. E.IGNALL, P.KOLESAR /74/ see(*)

. E.E.OSUNA, G.F.NEWELL /72/ "Control strategies for an Idealized Public Transportation system" Transportation science, vol 6 (57-72)

. J.TEGHEM Jr. /82/ "Optimal control of a ropeway" Technical report Department Gestion et Informatique - Faculté Polytechnique de Mons - Belgium

. *H.J.WEISS /81[b]/* "Further results on an infinite capacity shuttle with control at a single terminal" ORSA, vol 29,6 (1212-1217)

III.C. Clearing systems

. *S.NISHIMURA /79/* "Optimal interval for stochastic clearing systems" JORS Japan, vol 22,2 (95-104)

. *R.SERFOZO, Sh.STIDHAM Jr. /78/* "Semi stationary clearing processes" Stochastic Processes and their Applications, 6 (165-178)

. *Sh.STIDHAM Jr. /74/* "Stochastic clearing systems" Stochastic Processes and their Applications, 2 (85-113)

. *Sh.STIDHAM Jr. /77/* see(*)

. *W.WHITT /81/* "The stationary distribution of a stochastic clearing process" ORSA, vol 29,2 (294-308)

INSTRUCTIONS FOR THE AUTHORS

1. JORBEL accepts papers in the fields of Operations Research, Statistics and Computer Science.

2. Theoretical, applied and didactical papers, as well as documented computer programs are considered.

3. As far as possible, each issue will include a TUTORIAL PAPER. Authors are invited to submit also such papers giving didactical surveys on specific themes. These papers can be theoretical or applied. They should be comprehensible for non specialists and should establish the link between research and practice.

4. For the sake of effective communication it is recommended to the authors to have their papers written in *English.*

5. All submitted papers will be refereed. Only the authors will be responsible for their text.

6. As the journal is printed by offset, the manuscript should be camera ready, including figures and tables.

The *first page* is composed by the printer. It should only include:
- The title of the paper
- The names and affiliations of the authors (address included)
- An abstract of no more than 8 lines in English.

It is recommended to have the text of the manuscript typed double spaced with a prestige Elite type-head and the paragraph headings with a Script.
The formulas should also be typed and numbered on the right-hand side.
At each paragraph an 8 character jump should be observed.

7. The first author of published papers will receive 25 copies.

8. Papers should be submitted in two copies to one of the principal editors.

The authors should mention whether it is a tutorial paper.

ANNUAL SUBSCRIPTION RATE (4 issues)
BELGIUM: 600 BF. OTHER COUNTRIES: 700 BF

Sogesci: 53, rue de la Concorde, B-1050 Bruxelles, Belgium
B.V.W.B.: Eendrachtstraat 53, B-1050 Brussel, Belgium

Bank Account: 000-0027041-75

SOGESCI - B.V.W.B. BULLETIN

A Bulletin giving information on Operations Research, Statistics and Computer Science events is published quarterly by the SOGESCI-B.V.W.B.

Information on such events are to be mailed to G. Janssens:
RUCA, Middelheimlaan 1, B-2020 Antwerpen, Belgium