ETUDE DES EFFETS DE PERTURBATIONS CONNUES SUR DES SERIES SAISONNIERES

Roger ASTIER, Christian DUHAMEL

Laboratoire de Statistiques Appliquées
Université Paris-Sud
91405 ORSAY

Résumé: On étudie une série de données mensuelles de transport de voyageurs en isolant un comportement tendanciel à mois fixé et en évaluant les effets de certaines causes connues de perturbations de la série. Les résidus du modèle sont supposés auto-corrélés d'un mois sur l'autre et de variance périodique. Un exemple est présenté à partir de données fournies par la S.N.C.F. et un modèle de prévision est déduit. Le modèle général est linéaire avec résidus ARMA, les ruptures de pentes étant déterminées le plus souvent au jugé, et, en cas d'hésitation, pouvant l'être par moindres carrés résiduels.

PRESENTATION GENERALE

Le but de cette étude est, pour certaines séries mensuelles fournies par le ministère des transports, la conception de modèles de prévision à court terme prenant en compte des perturbations provoquées par les effets des grèves, des fêtes à date mobile, des augmentations tarifaires. Il s'agit aussi d'évaluer les effets statistiques sur les données mensuelles de ces perturbations. Les séries étudiées ont été : voyageurs-kilomètres du réseau principal S.N.C.F. et du trafic Air-Inter, consommation des véhicules à moteur, indice de circulation. On trouvera les résultats détaillés dans la référence (1) ; l'étude sur la série mensuelle "voyageurs-kms SNCF, réseau principal" est rapportée dans cet article.

Une méthode dite "Box et Jenkins avec fonction d'intervention" existe et a déjà été appliquée, à notre connaissance, à la série Air-Inter. Dans cette méthode, les saisonnalités ne sont pas évaluées et on suppose que de bonnes différenciations du type $(1-B)^d(1-B^{12})V_t$ permettront, partant de la série initiale V_t, de se ramener à une série stationnaire une fois éliminés les effets des perturbations. Ces derniers effets sont modélisés aux moyens de fonctions d'interventions (2,3). Outre le fait qu'on accroit ainsi la variance de la série à étudier et qu'on évalue ni tendance ni saisonnalité, on suppose implicitement qu'après élimination dans un modèle additif des composantes tendancielles et saisonnières et des perturbations, les résidus forment une série stationnaire. On élimine donc des comportements saisonniers des résidus eux-mêmes. Mais la variabilité résiduelle autour du comportement moyen n'a, a priori, aucune raison d'être la même en janvier et en Août.
Notre modèle propose une équation du type :

$$V_t = V_{(t)} + I_t + a_t$$

où V_t est la série brute des données ; $V_{(t)}$ est la série "tendance + saisonnalité". On

suppose une évolution linéaire par morceaux des composantes saisonnières à mois fixé. I_t est une fonction d'intervention prenant en compte les effets des perturbations au mois t, considérées ici comme variables exogènes. A un type de perturbation (ex. : grève importante) correspond une fonction d'intervention. Les effets perturbateurs intégrés dans I_t ont des causes connues : grèves, petites vacances dues à des "ponts", position dans l'année de fêtes mobiles, augmentation des tarifs des titres de transport. (a_t) est la série des résidus. On suppose que cette série (a_t) peut se ramener à une série stationnaire, donc à un modèle "Box et Jenkins", en corrigeant mensuellement : soit la variance des innovations (e_t) du modèle ARMA, c'est le cas du "Modèle 1" ; soit la variance des résidus initiaux (a_t), cas du "Modèle 2".

Une telle démarche suppose un calcul itératif qui a été rendu possible par l'utilisation du logiciel GENSTAT. Implanté notamment à ORSAY et au S.A.E. du ministère des transports, ce logiciel fournit un véritable langage très adapté aux traitements statistiques les plus divers. Avec, pour ce qui concerne cette étude : régression pondérée, estimation de modèles ARMA avec pondération sur les résidus, filtrage de séries chronologiques et estimation de données manquantes.

I - MODELE (1)

I-1 Equation générale du modèle

$V_t = m_t \beta + a_t$; V_t est l'observation au mois t

$\phi_p(B) a_t = \theta_q(B)\, w_t\, e_t$

m_t de dimension 1xl, intègre les évolutions tendancielles et saisonnières et les perturbations au mois t.

β de dimension lx1 est une vecteur colonne de paramètres

l est le nombre de variables et facteurs explicatifs

B est l'opérateur retard usuel : $B\, e_t = e_{t-1}$

ϕ_p et Θ_q sont des polynômes de degrés p et q

w_t est une fonction de période 12 introduite compte tenu de la différence des variances suivant le mois considéré.

(e_t) est un bruit blanc gaussien, centré, réduit.

La pondération w_t est apparue indispensable car il n'y a pas stationnarité d'un mois à un autre.

I-2 Choix du modèle et estimation des paramètres

Le choix des degrés p et q est fait à partir d'une première estimation du processus a_t par minimisation au sens des moindres carrés.

Les paramètres à estimer sont :

β : l valeurs ; généralement l est compris entre 20 et 30 environ

ϕ et θ : p+q valeurs ; généralement p+q assez faible ($\leqslant 5$)

w_t : 12 valeurs. On ne considère donc pas, à mois fixé, d'évolution de la variance.

On considère la vraisemblance $L(\beta, ,\theta,w)$. La détermination des paramètres par sa maximisation se fait en plusieurs étapes pour utiliser, autant que faire se peut, la résolution du modèle linéaire dans l'estimation de .

On procède donc comme suit :

(0) Une première estimation de β est réalisée sous l'hypothèse classique de blancheur du bruit a_t . Ceci permet en étudiant ce bruit de déterminer les degrés p et q de ϕ et θ et une première estimation de leurs coefficients et des variances $(w_t)_{t=1 \ldots 12}$: $\phi^{(0)}$, $\theta^{(0)}$, $w^{(0)}$.

(1) En fixant $\phi^{(0)}$, $\theta^{(0)}$, $w^{(0)}$, on calcule β qui maximise $L(\beta/\phi^{(0)},\theta^{(0)},w^{(0)})$. On trouve $\beta(1)$.

(2) Fixant $\beta^{(1)}$, on cherche ϕ et 0 maximisant $L(\phi,\theta/\beta^{(1)},w^{(0)})$. Dans cette étape on évalue $\theta^{(1)}$, $\phi^{(1)}$ et on estime $w_t e_t$: $w_t^{(1)} e_t$.

(3) Comme, par hypothèse, var e_t = 1, l'estimation de $w_t^{(1)} e_t$ donne la fonction de période 12 : $w_t^{(1)}$ variance empirique sur les résidus obtenus pour les mêmes mois.
On peut alors retourner au point (1) pour obtenir de nouvelles valeurs $\beta^{(2)}$, $\phi^{(2)}$, $\theta^{(2)}$, $w^{(2)}$ ... etc. On procède jusqu'à convergence des estimateurs. On trouvera ci-dessous des tableaux permettant de suivre l'évolution des estimations dans la suite des itérations. Dans certains cas, nous n'avions pas les renseignements nécessaires à la modélisation de perturbations importantes. Il apparait donc, au terme des estimations des paramètres, des valeurs aberrantes (2 en valeur absolue) pour certains résidus e_t. Dans ce cas, une nouvelle estimation est faite en supprimant les données brutes relatives à ces résidus (données considérées alors comme manquantes) et en recommençant une nouvelle estimation à partir des derniers résultats. La suppression d'une valeur aberrante ne se fait que s'il s'avère impossible d'en trouver la cause.

1-3 Modèlisation d'effets particuliers

Pour modèliser l'effet d'une grève par exemple, on introduit pour la période considérée une fonction d'intervention simple sous forme d'un facteur "grève" dont le niveau correspond au type de l'effet en question. Par exemple : 3 niveaux pour une grève : pas de grève, grève faible ou moyenne, forte. On tient compte éventuellement de la durée de l'effet dans le mois. Les paramètres à estimer sont entrés dans le vecteur β. Dans certains cas, celui des fêtes mobiles ou des augmentations de tarifs, nous avons dû procéder à une définition arbitraire des niveaux de facteur, dans l'impossibilité où nous sommes de multiplier outre mesure le nombre de paramètres à estimer. Les limites d'une telle modélisation tiennent au fait qu'une étude statistique ne peut se mettre en place qu'après une modèlisation subjective des effets connus. Ainsi les phénomènes d'évolution ne peuvent, à l'heure actuelle, être sérieusement entrés dans un modèle statistique. Il en est ainsi des comportements d'anticipation d'achat de titres de transports avant augmentation des tarifs pour lesquels les données à ce jour ne sont pas suffisantes. L'examen minutieux des résidus e_t du modèle considéré intégrant ces effets -mais non évolutifs- permet toutefois d'avoir une idée sur la liaison entre évolutions des comportements et variations du rythme des augmentations. La nécessité d'obtenir des estimateurs fiables (i.e. de faibles variances) interdit de tout prendre en compte et

nécessite des simplifications ou des regroupements qui ne valent que par la justesse de l'expérience que l'on peut avoir a priori sur le phénomène étudié.

1-4 Calculs et pratique

Pour effectuer les calculs, deux difficultés sont à résoudre :

. Sous la forme $V_t = m_t + a_t$ l'estimation par les moindres carrés de β faite par un programme de régression standard suppose que (a_t) est un bruit blanc, ce que nous excluons.

. Sous la forme $\phi(B)a_t = \theta(B)\ \xi_t$, les programmes usuels de maximisation de vraisemblance d'un processus ARMA supposent la stationnarité de ξ_t. Ce que nous excluons puisque les variables $\xi_t = w_t e_t$ ne sont pas de mêmes variances. Pour lever ces difficultés nous écrivons le modèle :

$$\phi(B)(V_t - m_t\ \beta) = \theta(B)\ w_t e_t$$

sous la forme suivante :

$$\frac{1}{w_t}\left\{\theta^{-1}(B)\phi(B)\ V_t - (\theta^{-1}(B)\phi(B)\ m_t)\ \beta\right\} = e_t$$

où e_t est un bruit blanc centré réduit.

β est alors estimé par la méthode des moindres carrés pondérés (par w_t) portant sur les variables :

$$\theta^{-1}(B)\phi(B)\ V_t \text{ et } \theta^{-1}(B)\phi(B)\ m_t$$

lesquelles sont obtenues à partir de m_t et des observations V_t par filtrage.

Puis lorsque β est estimé, on calcule $V_t - m_t \cdot \beta$

II - APPLICATION A L'ETUDE DE LA SERIE "SNCF - GRANDES LIGNES"

La série présentée est celle du nombre mensuel de voyageurs-kilomètres du réseau grandes lignes de la SNCF. Les données vont de janvier 1969 à septembre 1982 et sont obtenues à partir des ventes de titres de transports effectuées dans le mois. Elle ne correspondent donc pas exactement aux voyages effectifs dans le mois.

II-1 Informations intégrées dans le modèle

Ces informations entrées dans le vecteur m_t pour chaque mois t sont les suivantes :

a) Evolution tendancielle

Une rupture de pente est localisée à partir de janvier 1976 et correspond à toutes les études de désaisonnalisation effectuées sur cette série. D'autre part, des études préliminaires montrent qu'aussi bien avant qu'après cette rupture, les pentes des mois suivants : janvier, février, mars, avril, mai, ne sont pas significativement différentes. Il en

est de même pour octobre, novembre, décembre. On regroupe donc ces mois entre eux pour l'estimation de la pente, et il reste 12 paramètres de pentes : 6 avant rupture et 6 après. Appelons m(t) le nom du mois t. Ainsi m(1) = m(13) = janvier. L'évolution tendancielle de t = 1 (janvier 69) à t = 145 (septembre 82) s'écrit donc $t \to T_t$ avec : jusqu'en décembre 1975 (t=84) : $T_t = a_{m(t)} + b^{(1)}_{m(t)} t$. Et à partir de janvier 1976 (t=85) :

$$T_t = a_{m(t)} + b^{(1)}_{m(t)} \cdot 85 + b^{(2)}_{m(t)} \cdot (t-85).$$

b) Prise en compte des grèves

A partir du classement fourni par la SNCF et après une première étude statistique, les jours de grève dans le mois sont classés pour cette étude en deux types : "faible ou moyen" et "fort". Rappelons que les données ont trait au nombre de billets vendus dans le mois et non au nombre de voyageurs effectivement transportés. Des grèves en début de mois ou en fin de mois n'ont donc pas le même effet sur la donnée disponible. Ceci explique en partie l'importance des écarts-types pour les estimations de ces effets. On cherche alors à estimer le nombre de voyageurs x kilomètres, (comptés à partir des ventes mensuelles de billets), perdu par jour de grève de type "fort" ou "moyen faible".

c)Prise en compte des augmentations tarifaires

Une augmentation des tarifs de transport à la SNCF entraîne généralement, dans la période qui précède cette augmentation, des anticipations d'achat des billets. Pour le mois de la hausse des tarifs et éventuellement pour le mois suivant, cela entraîne une baisse des ventes comptabilisées. On suppose que ces hausses n'ont pas d'effets à long terme. Ce qui est sans doute faux, il y a vraisemblablement une évolution des comportements qui montre une accélération de ce phénomène d'anticipation. Accélération difficile à entrer dans un modèle statistique vu le nombre d'occurences mais dont on peut se faire une idée en étudiant les résidus du modèle pour les mois liés à ces hausses de tarifs.
Nous introduisons un facteur à trois niveaux "AUGmentation" pour le mois t : Supposons que l'augmentation ait lieu au mois t_o. Alors $AUG_{t_o-1}(1) = 1$, $AUG_{t_o}(2) = 1$, $AUG_{t_o+1}(3) = 1$.
Dans les autres cas, $AUG_t(i) = \theta, i = 1, 2, 3$. Et la régression est faite sur ce facteur.

Les dates et pourcentages d'augmentation sont les suivants :

1er octobre 69 : 5,6%
4 janvier 71 : 5,1%
16 mai 73 : 5,33 %
15 avril 75 : 8,5 %
18 avril 77 : 6,5 %
1er février 79 : 7,6%
17 mars 80 : 9,4 %
1er septembre 81 : 10,2 %

5 janvier 70 : 5,5%
1er mai 72 : 5%
1er avril 74 : 7,6%
6 janvier 76 : 7,8%
1er mai 78 : 15,7%
1er septembre 79 : 5%
30 mars 81 : 7,4 %
1er mars 82 : 11,2%

Si les augmentations se situent généralement en début de mois (le 30 mars 81 est assimilé au 1er avril 81), celles qui se situent dans la deuxième quinzaine posent problème dans la mesure où s'est perdue l'information sur les dates auxquelles elles étaient annoncées, dates importantes pour l'achat des billets. Aussi nous avons tenu compte, pour ces mois-ci, d'un examen des résidus fournis par une analyse faite sans

considérer le facteur "augmentation tarifaire", pour attribuer les niveaux de ce facteur. Ainsi en mars 1980 (augmentation le 17 du mois), les anticipations d'achat de la première quinzaine du mois semblent avoir compensé la baisse d'après le 17. Nous avons donc pris :

$AUG_{février\ 80}(i) = AUG_{mars\ 80}(i) = O$ pour i = 1, 2, 3, $AUG_{avril\ 80}(3)=1$

d) Existence de fêtes mobiles

Il s'agit essentiellement de Pâques et de la Pentecôte. Les congés de Pâques se trouvent en mars, en avril ou à cheval sur les deux mois. Pentecôte est en mai ou juin. Nous avons également estimé l'effet d'un "pont" type Toussaint, premier mai, 14 juillet, 15 août...sur les ventes de billets dans le mois en introduisant un facteur dont le niveau dépendait de la place du jour férié dans la semaine. Mais aucun résultat significatif n'est apparu permettant de conclure sur l'existence d'un "effet pont". Ce résultat semble en accord avec ce qui était connu : si un effet "pont" peut apparaître lorsqu'on considère des données mensuelles ou journalières, il devient statistiquement non significatif sur des données mensuelles. Soit parce que la variabilité de ces données absorbe l'effet, soit parce qu'une augmentation des voyages durant ces quelques jours entraîne une diminution sur les autres jours du mois, ou l'inverse.

Enfin, des corrections faites pour tenir compte du nombre de jours ouvrés, ou, ce qui revient au même, du nombre de week-ends dans le mois, n'ont apporté aucune amélioration significative des résultats. Se ramener à un trafic moyen par jour permettrait d'avoir des coefficients constants de l'évolution tendancielle comparables ; mais ces coefficients sont de peu d'intérêt, les coefficients importants étant les pentes de cette évolution, calculées pour chaque mois. Nous nous en sommes tenus aux données brutes V_t, ce qui rend plus aisée l'interprétation des résultats, notamment la perte par jour de grève.

II-2 Equation détaillée du modèle de régression

V_t = nombre de voyageurs x kilomètres le mois t, comptabilisés à partir des billets vendus le mois t.

. t = 1 pour janvier 69, t = 84 pour décembre 75

. t = 145 pour septembre 82

. t = a + 12 * m(t) où m(t) représente le mois de 1 à 12 et "a" l'année.

$$V_t = \alpha_{m(t)} + \beta_{m(t)} \cdot t + \sum_{i=1}^{2} \gamma_i\, g_t(i) + \sum_{j=1}^{3} \zeta_j f_t(j) + \sum_{k=1}^{3} \pi_k\, AUG_t(k) + \text{résidu } a_t$$

$\alpha_{m(t)}$ = niveau de base (12 paramètres) du mois m(t).

$\beta_{m(t)}$ = pente associée au mois m(t) : voir ci-dessus.

(6 x 2 = 12 paramètres).

$g_t(1)$ indique le nombre de jours de grève "forte" pendant le mois t.

$g_t(2)$ pour le nombre de jours de grève "moyenne" ou faible" au mois t. Donc 2 paramètres γ_1 et γ_2. S'il n'y a pas eu de grève le mois t : $g_t(1) = g_t(2) = O$.

$f_t(j)$ pour j = 1, 2, 3 prend en compte un effet possible Fêtes de Pâques et de Pentecôte. Les variables prennent les valeurs 0 ou 1 suivant :

$f_t(1)$ = 1 si le congé de Pâques (à Paris) est à cheval sur les mois t et t+1.

$f_t(2)$ = 1 si ce congé est uniquement sur le mois t.

$f_t(3)$ = si le congé de Pentecôte est le mois t.

Donc 3 paramètres à estimer ($\delta_1, \delta_2, \delta_3$)

$AUG_t(k)$, k = 1, 2, 3 prend en compte sur le mois t les effets des augmentations tarifaires au cours du mois t-1, t ou t+1.

$AUG_t(1)$ = 1 si l'augmentation a lieu au mois t+1.

$AUG_t(2)$ = 1 si l'augmentation a lieu au mois t.

$AUG_t(3)$ = 1 si l'augmentation a lieu au mois t-1.

(voir ci-dessus dans le cas d'augmentation en fin de mois).
On a donc les 3 paramètres π_k à estimer.

Les résidus (a_t) sont modèlisés par une équation du type ARMA(p, q) :

$\phi_p(B)a_t = \Theta_q(B)w_t.e_t$. Où la fonction $t \rightarrow W_t$ est de période 12 et prend en compte les différences sur les écarts types des innovations. Soit p+q+12 paramètres.

Résultats des calculs effectués

Une première régression est faite en minimisant Σa_t^2, c'est-à-dire sans modéliser les résidus (a_t). Une étude des auto-corrélations des résidus est faite, qui peut sembler assez décevante, vu le but de l'étude : les auto-corrélations ne sont guère significatives. Mais cela n'a pas grand sens, car les variances des a_t, calculées à mois fixé, sont très différentes. On modèlise donc (a_t) comme un AR(2) avec pondération sur les innovations e_t :

$$(1 - \phi_1 B - \phi_2 B^2)\, a_t = W_t e_t$$

et les itérations se succèdent comme il a été décrit précédemment. La notion de variance résiduelle n'a pas ici le sens habituel puisque var e_t = 1 et la qualité de la prévision est donnée par $W_t e_t$. Le modèle AR(2) sur les résidus donne ϕ_1 = 0.39 (0,06) et ϕ_2 = O.10 (0.06). Entre parenthèses figure l'écart-type de l'estimateur.
On obtient $\frac{1}{n} \Sigma_t W_t^2 e_t^2$ = 22 460. Pour les auto-corrélations de la série (e_t), on trouve $\sum_{i=1}^{38} \hat{\rho}_i^2$ = 42.7 comme valeur, sous l'hypothèse de blancheur du bruit, d'un CHI2 à 36 degrés de liberté. L'autocorrélation $\hat{\rho}_6$ et partielle $\hat{\pi}_6$ sont significatives. On considère donc un deuxième modèle type ARMA pour les résidus a_t :

$$(1-\phi_1 B-\phi_2 B^2)a_t = (1-\theta_6 B^6)\, W_t e_t.$$

Dans ce modèle on obtient : $\frac{1}{n}\Sigma w_t^2 e_t^2 = 11524$ donc un gain appréciable pour l'ajustement du modèle et, une valeur du CHI2 à 36 degrés de liberté de 36.8. On a alors :

$$\phi_1 = 0.27\ (0.06), \phi_2 = 0.03\ (0.06), \hat{\theta}_6 = -0.35\ (0.06).$$

On a par ailleurs des variations sur l'estimation des coefficients β. Les tableaux suivants résument les résultats.

Tableau T1 : estimation des coefficients β dans 3 cas :

C1 - estimation classique par minimisation de Σa_t^2

C2 - estimation générale pondérée sur les résidus avec $(1-\phi_1 B-\phi_2 B^2)a_t = W_t e_t$

C3 - estimation générale pondérée sur les résidus avec $(1-{}_1B-{}_2B^2)a_t = (1-\theta B^6)W_t e_t$

C4 - (voir ci-dessous : modèle 2). Modèle ARMA sur les résidus normalisés :

$$(1-\phi_1 B-\phi_2 B^2)\frac{a_t}{W_t} = (1-0\,B^6)\, e_t\ .$$

On voit que si les variations sur les estimations des coefficients de tendance et saisonnalité sont assez faibles, il n'en est pas de même des estimations des effets des perturbations, sauf en ce qui concerne l'effet des grèves, la perte par jour de grève "forte" peut être estimée à environ 30 unités de voyageurs x kms (1,5 % de la valeur mensuelle) et par jour de grève qualifiée de "moyenne" ou "faible", à 22. Le comportement décrit pour les anticipations d'achat reste significatif dans tous les cas mais les résultats sont instables quand on passe d'un modèle à un autre. Le phénomène est en pleine évolution à la fois par le rythme des augmentations de tarif et par le comportement des usagers. La prise en compte de corrélations d'ordres voisins de 6 dans le modèle $\phi_2(B)a_t = (1-\theta_6 B^6)\, W_t e_t$ peut biaiser le résultat, dans la mesure où les périodes entre deux augmentations tarifaires sont généralement voisines de 6 ou 12 mois. On peut toutefois considérer que le mois précédent une augmentation, on a un accroissement des ventes de billets d'environ 150 unités de voyageurs x km (écart-type de 40), et, durant le mois de l'augmentation et le mois suivant des baisses respectives d'environ 120 (35) et 70 (30). Ceci semblerait indiquer que les anticipations d'achat ne compensent pas exactement les pertes occasionnées par une augmentation des tarifs, mais cette conclusion n'est pas statistiquement significative. Enfin, le fait de placer les vacances scolaires de Pâques (région parisienne) en Avril essentiellement implique pour ce mois une augmentation de 230 voyageurs x km (100). On a aussi bien sûr un résultat du même ordre lorsque les vacances de Pâques se situent à cheval sur Mars/Avril, mais ce résultat n'est pas significatif et difficile à chiffrer. Il est de toute façon inférieur d'environ la moitié au résultat précédent. De même on a, pour le mois contenant le congé de Pentecôte une augmentation non significative du nombre de billets vendus dans le mois. Le tableau T2 donne l'évolution des estimations des coefficients des paramètres ARMA et des variances W_t dans les trois cas C2, C3, C4 évoqués ci-dessus. La variation des

estimations des coefficients ARMA est importante, mais celle des variances résiduelles l'est beaucoup moins (environ 10%) et assez lente. On voit toutefois que le fait de ne pas considérer de différences suivant les mois entre les variances résiduelles introduit des erreurs importantes sur les autocorrélations et donc sur la modèlisation des résidus. Le tableau T3 donne les ajustements des données au mois t de Octobre 1980 à Septembre 1982 inclus ; ajustements obtenus en annulant les innovations pour chaque mois. On obtient donc une estimation a posteriori de la prévision $E(V_t / V_{t-1}, V_{t-2}, \ldots)$, mais à partir d'un modèle estimé sur l'ensemble des données, y compris V_t. Les dernières valeurs sont relatives aux prévisions pour Octobre, Novembre et Décembre 1982, à l'aide de données disponibles jusqu'en Septembre 1982 inclus. On peut aussi estimer une ou des valeurs manquantes qui sont considées comme autant de paramètres en plus à estimer par moindres carrés, l'ensemble des paramètres (β, o, O, (W_t), valeurs manquantes) est estimé dans un seul modèle en minimisant $\sum_t a_t^2$. Cela permet dans le cas où un résidu $|e_t|$ est anormalement grand, c'est-à-dire supérieur à deux fois l'écart-type $\hat{\sigma}_e$ de supprimer la donnée V_t correspondante et d'estimer l'ensemble des paramètres y compris cette donnée manquante. Dans cette série SNCF aucune valeur e_t ne pouvait réellement être considérée comme aberrante. Dans le tableau T4, six données ont été supprimées simultanément puis réestimées suivant cette procédure. En déclarant manquantes les données qui suivent les données disponibles, ou en utilisant les estimations faites (le calcul peut se faire "à la main"), on construit des valeurs pour la prévision des mois à venir. Le tableau T5 donne les prévisions pour les mois d'Octobre, Novembre, Décembre 1982 suivant chacun des modèles considérés et isole la partie $m_t \beta$ de la composante a_t.

III - MODELES DU TYPE (II)

Le modèle étudié précédemment suppose que les résidus (a_t) de l'équation $V_t = m_t \beta + a_t$ satisfont à l'équation $\phi_p(B) a_t = \Theta_q(B) W_t e_t$, où (e_t) est un bruit blanc gaussien. Après inversion du polynôme $o_p(B)$, on peut aussi écrire, avec des constantes ψ_i :

$$a_t = \Phi_p^{-1}(B)\, \Theta_q(B)\, W_t e_t = W_t e_t + \sum_{i>1} \psi_i\, W_{t-i}\, e_{t-i}$$

$W_t\, e_t$ représente la part d'erreur du modèle dans la prévision faite au mois t-1 pour le mois t. L'imprévu $W_{t-i}\, e_{t-i}$ du mois t-i se retrouve avec le poids ψ_i et le signe de ψ_i dans le résidu a_t du mois t .

Dans un deuxième modèle que nous désignerons comme le "modèle II", nous proposons un traitement plus classique : les résidus normalisés de la régression sont supposés suivre un modèle ARMA usuel. L'équation générale est donc :

$$V_t = m_t \beta + a_t \qquad , \text{var } a_t = W_t^2$$

$$\Phi_p(B)\, \frac{a_t}{W_t} = \theta_q(B)\, e_t$$

Les notations et les variables de régression sont les mêmes que pour le modèle (I).
Le modèle s'écrit également :

$$V_t = m_t\beta + W_t\Phi_p^{-1}(B)\,\Theta_q(B)\,e_t$$

L'estimation des coefficients est simplifiée puisqu'on n'a plus à considérer l'étape de regression pondérée du premier modèle. En effet, la minimisation de $\Sigma\, e_t^2$ se fait en suivant la procédure :

(1) $V_t = m_t\beta + a_t$ minimiser $\Sigma a_t^2 \Rightarrow \beta^{(0)}$

(2) $a_t^{(0)} = V_t - m_t\beta^{(0)} \Rightarrow W_t^{(0)}$, variance périodique

(3) Etude de la série $(a_t^{(0)}/W_t^{(0)})$, identification de (p,q), estimation des polynômes du modèle ARMA $\Rightarrow \Phi_p^{(0)}$, $\Theta_q^{(0)}$.

(4) $V_t - m_t = W_t\Phi_p^{-1}(B)\,\theta_q(B)e_t$ s'écrit :

$$\theta_q^{(0)-1}(B)\phi_p^{(0)}(B)\,\frac{1}{W_t^{(0)}}\,V_t - \left[\theta_q^{(0)-1}(B)\,\phi_p^{(0)}(B)\,\frac{m_t}{W_t}\right]\beta = e_t .$$

Minimiser $\Sigma\, e_t^2 \Rightarrow \beta^{(1)}$

(5) Poser $\beta^{(0)} = \beta^{(1)}$ et retourner en (2).

Interprétation du modèle II

Les résidus (a_t) s'écrivent :

$$a_t = W_t\Phi^{-1}(B)\,0(B)\,e_t = \sum_{i\geq 1}\psi_i\,W_t\,e_{t-i} + W_t\,e_t$$

$W_t e_t$ représente toujours la part d'erreur dans une prévision faite au mois t-1 pour le mois t. On suppose donc que la quantité $W_{t-i}\,e_{t-i}$, écart à la prévision du mois t-i, se retrouve au cours du mois t avec le facteur et avec le même signe que $\psi_i\,W_t/W_{t-i}$.

Les résultats dans ce modèle sont portés parallèlement à ceux du modèle (I) précédent. Une première étude avec un modèle AR(2) :

$$(1-\phi_1 B-\phi_2 B^2)\,a_t/W_t = e_t$$

donne $\phi_1 = 0.30\ (0.008)$, $\phi_2 = 0.16\ (0.07)$ et une variance résiduelle de 0.80. Le CHI2 du

test de blancheur donne pour 36 degrés de liberté la valeur 37 et laisse sur le bruit (e_t) une auto-corrélation significative à l'ordre 6.

Le modèle $(1-\phi_1 B-\phi_2 B^2)\, a_t/W_t = (1-\Theta B^6)e_t$ donne après estimation :

$$\phi_1 = 0.28\ (0.08), \phi_2 = 0.03\ (0.08),\ \theta = -0.48\ (0.07)$$

et une variance des e_t de 0.67. Le CHI2 pour 36 degrés de liberté est de 38. Calculées pour chaque mois, les variances résiduelles des e_t calculées sur 14 années prennent les valeurs suivantes qui ne permettent pas de rejeter l'hypothèse d'une variance des innovations constante :

Janv : 0.88 Février : 0.64 Mars : 0.82 Avril : 0.80 Mai : 0.45 Juin : 0.59 Juil : 0.97
Août : 0.53 Sept : 0.54 Oct : 0.75 Nov : 0.46 Déc : 1.04

Tableau T1 : Estimation du vecteur β

Niveau Janvier	2322	(47)	2257	(87)	2230	(63)	2238	(60)
Février	2163	(47)	2079	(83)	2067	(67)	2076	(66)
Mars	2470	(65)	2458	(107)	2476	(90)	2496	(96)
Avril	2285	(87)	2276	(129)	2282	(105)	2312	(112)
Mai	2597	(60)	2541	(82)	2529	(60)	2540	(56)
Juin	2989	(90)	3010	(82)	2996	(68)	3025	(62)
Juillet	3290	(87)	3293	(91)	3234	(90)	3220	(87)
Août	3242	(89)	3240	(56)	3147	(63)	3142	(63)
Septembre	2472	(95)	2455	(63)	2376	(65)	2394	(59)
Octobre	2349	(61)	2370	(47)	2363	(48)	2394	(43)
Novembre	2156	(62)	2168	(46)	2172	(44)	2197	(43)
Décembre	2829	(64)	2856	(55)	2848	(54)	2875	(51)
Pentes avant 1976								
Janvier à mai	12.0	(0.6)	13.5	(1.3)	13,3	(0.9)	13.3	(0.8)
Juin	12.8	(1.5)	12.6	(1.4)	13.0	(1.1)	12.7	(1.)
Juillet	12.1	(1.5)	12.0	(1.5)	12.9	(1.5)	13.2	(1.4)
Août	5.5	(1.5)	5.5	(1.)	6.4	(1.0)	6.6	(1.)
Septembre	12.1	(1.6)	12.2	(1.0)	13.4	(1.0)	13.3	(0.9)
Octobre à Décembre	13.1	(0.9)	12.8	(0.8)	12.9	(0.7)	12.6	(0.7)
Pentes après 1975								
Janvier à Mai	6.1	(0.7)	5.3	(1.3)	6.2	(0.9)	6.0	(0.9)
Juin	6.9	(1.5)	6.4	(1.4)	6.1	(1.1)	5.9	(1.)
Juillet	4.5	(1.5)	4.4	(1.5)	4.5	(1.5)	4.2	(1.4)
Août	3.1	(1.5)	3.5	(0.9)	4.3	(1.0)	4.1	(1.0)
Septembre	3.8	(1.5)	4.5	(1.)	3.7	(1.0)	3.6	(0.9)
Octobre à Décembre	8.4	(1.0)	8.6	(0.8)	8.06	(0.8)	8.0	(0.8)
Pâques Mars/Avril	162	(70)	106	(95.)	49.3	(86)	27.7	(100)
Pâques Avril	264	(96)	258	(113)	226.	(96)	208.	(106)
Pentecôte Mai ou Juin	46.	(49)	53	(32)	25	(29)	24.	(28)
Augmentations								
mois précédant	174.	(40)	114	(36.)	170	(37)	164.	(36)
mois	-140.	(39.)	-214	(37)	-121	(32)	-139	(32.)
mois suivant	-36	(37.)	-74	(29)	-54	(26)	-68	(26)
Par jour de grève								
Forte	-32	(6.3)	-30	(4.4)	-27	(4.5)	-28	(4.3)
Moyenne-faible	-24	(5.5)	-21	(3.2)	-22	(3.3)	-24	(3.2)
	Régression "classique"		$(1-\phi_1 B-\phi_2 B^2)a_t = W_t e_t$		$(1-\phi_1 B-\phi_2 B^2)a_t = (1-\theta_6 B^6)W_t e_t$		$(1-\phi_1 B-\phi_2 B^2)\frac{a_t}{W_t} = (1-\theta_6 B^6)\, e_t$	

Modèle sur les résidus	ϕ_1	ϕ_2	θ_6	JAN W1	FEV W2	MARS W3	AVR W4	MAI W5	JUIN W6	JUIL W7	AOUT W8	SEPT W9	OCT W10	NOV W11	DEC W12
	O.	O.		121	123	173	145	93	93	116	69	78	72	59	110
Cas C2	O.19	O.14		123	118	168	143	89	94	111	66	78	68	59	108
$(1-\phi_1B-\phi_2B^2)a_t = W_t e_t$	O.36	O.09		150	120	168	164	84	91	108	54	67	60	57	110
	O.29	O.08		156	119	169	167	84	91	108	52	66	58	57	110
Cas C3	O.15	O.03	-O.17	108	129	176	140	86	89	118	74	78	75	52	111
$(1-\phi_1B-\phi_2B^2)a_t =$	O.23	O.04	-O.38	106	137	167	137	67	86	119	71	73	66	49	113
$(1-\theta_6B^6)W_t\ e_t$	O.27	O.03	-O.35	110	143	168	142	65	88	117	67	70	63	50	112
	(O.06)	(O.06)	(O.06)												
Cas C4	O.23	O.08	-O.18	122	145	180	127	83	97	127	90	87	72	68	109
$(1-\phi_1B-\phi_2B^2)a_t$	O.25	O.03	-O.32	125	155	184	129	79	99	128	95	88	69	70	111
$= (1-\theta_6B^6)e_t$	O.28	O.04	-O.48	123	159	189	131	79	101	130	98	89	68	71	112
	(O.07)	(O.07)	(O.07)												

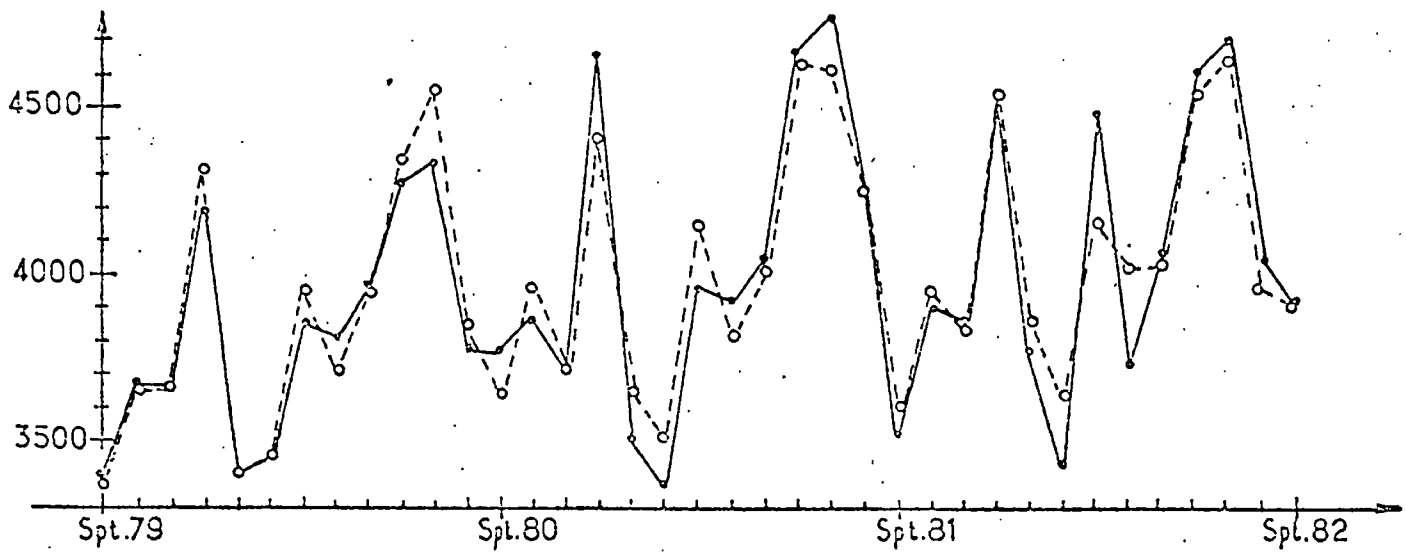
Tableau T2 : Evolution rapportée toutes les deux itérations des coefficients , θ , W.

Tableau T3 : Données ajustées $\hat{E}(V_t/V_{t-1}, V_{t-2}, \ldots)$

Quatre cas sont considérés : C2 : $\phi_2(B)a_t = W_t e_t$ C3 : $\phi_2(B)a_t = (1-\theta B^6)W_t e_t$

C4 : $\phi_2(B)a_t/W_t = e_t$ C5 : $\phi_2(B)a_t/W_t = (1-\theta B^6)e_t$

t	Données réelles	C2	C3	C4	C5
Avr. 81	3904	3823	3815	3805	3772
mai	4055	3970	4003	3986	4008
juin	4670	4563	4634	4609	4664
juil	4782	4608	4617	4667	4613
août	4169	4068	4146	4125	4159
sept	3520	3603	3595	3654	3575
oct	3910	4070	3957	3997	3960
nov	3855	3930	3833	3847	3835
dec	4514	4606	4534	4540	4526
Jan 82	3774	3799	3865	3790	3894
févr	3440	3630	3650	3623	3649
mars	4487	4138	4152	4089	4105
avril	3722	3891	4029	3972	3960
mai	4067	4097	4047	4104	4064
juin	4622	4583	4551	4551	4538
juill	4707	4660	4664	4657	4624
août	4055	3972	3969	3997	3971
sept 82	3916	3824	3939	3861	3911



Données mensuelles SNCF en millons de voyageurs-km: données réelles

données ajustées (modèle C3)

$\hat{E}(V_t / V_{t-1}, V_{t-2}, \ldots)$ pour t de septembre 1979 à septembre 1982.

Tableau T4 : Réestimation des données déclarées simultanément manquantes
Modèle : $\phi_2(B)\, a_t = W_t\, e_t$.

date	MAI 71	DEC 72	AOUT 77	SEPT 77	OCT 80	NOV 80
donnée réelle	2872	3442	3708	3504	3870	3724
réestimation	2967	3406	3776	3613	3992	3771

Tableau T5 : Prévision suivant chaque modèle. Dernière donnée disponible : Septembre 1982. Entre parenthèse la composante résiduelle $\hat{a}_t$

Modèle	$\phi_2(B)a_t = W_t\, e_t$	$\phi_2(B)a_t = (1-\theta B^6)W_t e_t$	$\phi_2(B)\frac{a_t}{W_t} = e_t$	$\phi_2(B)\frac{a_t}{W_t} = (1-\theta B^6)e_t$
oct. 82	4179 (20)	4033 (-78)	4216 (32)	4062 (-42)
Nov. 82	3973 (8)	3906 (-23)	3996 (15)	3905 (-11)
Déc. 82	4663 (1)	4640 (27)	4678 (5)	4648 (48)

Données réelles : octobre : 4192, novembre : 3621, décembre : 4560

IV - AUTRES SERIES

D'autres séries mensuelles ont été étudiées par la même méthode, notamment la série AIR INTER (en millions de passager-kms) et l'indice de consommation des véhicules à moteur obtenu à partir des ventes mensuelles aux pompistes, donc perturbé par des phénomènes de stockage, commande avant hausse, rétention de stocks. De tous ces phénomènes nous n'avons pas pu tenir compte faute d'informations précises. Il apparaît des données "aberrantes" sur les séries des innovations. Les paramètres sont ré-estimés après suppression de ces données.

IV-1 Série Air Inter

N'ayant pu obtenir pour la période étudiée (janvier 70 à octobre 81) les informations précises sur les grèves et leur importance, nous avons introduit comme seules perturbations les congés de Pâques et Pentecôte. La partie régression linéaire considère une pente égale pour chaque mois, avec rupture en octobre 1977 :

$$V_{(t)} = c_{m(t)} + b_1 t \text{ , si } t \leq 93 \text{ (octobre 1977)}$$

$$V_{(t)} = c_{m(t)} + b_1 93 + b_2(t-93) \text{ , si } t > 93$$

Les résidus a_t sont modélisés par une équation AR(3) :

$$(1 - \phi_1 B - \phi_2 B^2 - \phi_3 B^3)a_t = w_t \cdot e_t$$

Il apparait après examen des innovations sept valeurs aberrantes, c'est-à-dire supérieures à 2 en valeurs absolues. Dans le tableau des résultats qui suit, "C_1" correspond à l'estimation classique au sens des moindres carrés sur les résidus (a_t) ; "C_2" à la régression pondérée sur résidus AR(3) ; "C_3" est comme "C_2", mais en excluant les valeurs aberrantes ; et "C_4" est la régression non pondérée sur résidus AR(3), valeurs aberrantes exclues.

avant oct. 77	1.41 (0.07)	1.48 (0.07)	1.44 (0.07)	1.45
après oct. 77	3.22 (.15)	3.00 (.15)	3.28 (0.15)	3.20 (0.17)
Pâques MARS/AVRIL	40.49 (12.61)	38.01 (14.39)	52.19 (11.61)	26.44 (8.64)
Pâques AVRIL	44.34 (17.80)	40.93 (13.69)	48.03 (11.94)	27.17 (9.68)
Pentecôte MAI OU JUIN	14.3 (8.46)	-10.56 (3.85)	-5.27 (3.14)	-6.28 (4.47)
	estimation au sens des moindres carrés	régression pondérée sur résidus ARMA	régression pondérée sur résidus ARMA, valeurs aberrantes exclues	régression non pond. sur résidus ARMA Valeurs aberrantes exclues
	C1	C2	C3	C4

On obtient également pour les coefficients ϕ du polynôme d'autorégression et les écarts-types w_i des innovations les valeurs suivantes :

Cas	ϕ_1	ϕ_2	ϕ_3	w_1	w_2	w_3	w_4	w_5	w_6	w_7	w_8	w_9	w_{10}	w_{11}	w_{12}
C_2	.34	-.11	.18	6.2	14.9	23.1	20.2	7.22	17.3	11.1	21.3	16.7	12.5	37.0	14.1
C_3	.16	.04	.30	6.5	14.5	24.6	9.3	8.3	7.2	12.8	17.4	8.8	12.9	13.3	8.1

On voit l'importance de l'exclusion des valeurs aberrantes et de la pondération des innovations sur l'estimation à la fois des coefficients et des paramètres de perturbation.

IV-2 Série consommation de carburants (essence + super)

Les données varient entre 1000 et 1900 pour la période étudiée de janvier 1970 à novembre 1981. Pour étudier l'évolution tendancielle et saisonnière, le modèle est le même que pour Air Inter. Les perturbations considérées sont les grèves SNCF, les positions de Pâques et de Pentecôte. Les résidus a_t sont modélisés par :

$$(1-\phi_1 B-B-\phi_3 B^3-\phi_6 B^6)a_t = w_t e_t$$

Les résultats suivants montrent un faible effet significatif pour les grèves de transports en commun, la différence entre grève forte et faible n'étant pas significative. La position en avance de Pâques entraîne une diminution de la donnée de mars de 5 à 8% et la présence de la Pentecôte dans le mois une augmentation de l'ordre de 5%.

	Estimation au sens des moindres carrés	Régression pondérée sur résidus ARMA	Régression pondérée sur résidus ARMA valeurs aberrantes exclues	Régres.non pond. avec résidus ARMA val. aberrantes exclues
Pente avant 04/74	$6.72_{(.51)}$	$6.59_{(.82)}$	$6.64_{(.64)}$	$6.35_{(1.20)}$
Pente avant O4/74	$2.36_{(.22)}$	$2.46_{(.21)}$	$2.42_{(.19)}$	$2.23_{(0.33)}$
Par jour grève forte	$6.22_{(3.7)}$	$5.45_{(1.2)}$	$3.83_{(1.2}$	$4.70_{(2.6)}$
Par jour grève faible	$4.84_{(3.0)}$	$6.73_{(1.7)}$	$6.46_{(1.5)}$	$4.75_{(2.0)}$
Pâques sur MARS/AVRIL	$-47.9_{(41)}$	$-105_{(30)}$	$-77_{(26)}$	$-57_{(33)}$
Pâques sur AVRIL	$78.10_{(59)}$	$-56_{(47)}$	$-26_{(43)}$	$17_{(48)}$
Pentecôte	$39.98_{(28)}$	$77_{(14)}$	$60_{(14)}$	$50_{(24)}$

V - CONCLUSION

Nous avons introduit un modèle pour estimer simultanément l'évolution tendancielle et saisonnière ainsi que les effets de perturbations connues sur une série mensuelle. Pour les résidus, supposés autocorrélés, il y a possibilité de différences entre variances à mois fixé. L'hypothèse d'évolution linéaire par morceaux de la tendance à mois fixé peut sembler arbitraire, mais semble bien s'ajuster aux séries considérées. Envisager d'autres types d'évolution tendancielle que celles entrant dans le modèle linéaire compliquerait beaucoup les calculs. Ne pas se poser le problème de l'estimation de cette évolution revient à construire une analyse des interventions dans un modèle SARIMA de Box et Jenkins. Mais cela implique l'hypothèse d'équi-variabilité de tous les mois et on a vu que sur les données traitées, c'était difficilement acceptable.

Pour estimer les effets de perturbations, sont introduites des fonctions d'intervention du type $a1_{t_o}(t)$ dans le cas le plus simple, ou $a1_{t_o}(t) + b1_{t_o}(t) + c1_{t_o+1}(t)$ dans le cas des augmentations tarifaires. Dans l'impossibilité de trop augmenter le nombre de paramètres à estimer (ici : a, b, c), ces effets sont, pour une perturbation donnée, supposés constants sur l'ensemble de la série étudiée et les mêmes pour tous les mois. En fait, l'examen des résidus et des innovations montrent que les phénomènes d'anticipation à la hausse, par exemple, sont en pleine évolution, comme l'a encore montré la récente augmentation tarifaire du printemps 1983. Notons que cet effet dépend beaucoup des dates auxquelles sont prévisibles les augmentations. A moins de se plonger dans les conjectures de la presse écrite et parlée de l'époque, le statisticien ne dispose que des données mensuelles brutes, de la date et du pourcentage d'augmentation. On pourrait bien sûr faire dépendre les coefficients du temps en écrivant :

$$a = a(t_o) = a' + a''t_o.$$

De nombreux effets existant réellement, comme ceux des "ponts", peuvent disparaître lorsque les données sont exprimées mensuellement. Aussi, la conclusion n'est jamais : "cet effet n'existe pas", mais :"cet effet n'apparait pas de façon significative dans les données mensuelles". Enfin, les congés de Pâques ont été considérés de façon simplificatrice en observant les dates des congés scolaires de la seule région parisienne, la plus importante pour les migrations de cette époque.

Les deux modèles considérés pour modéliser les résidus - variance périodique pour les innovations (modèle 1) ou pour les résidus (modèle 2) - ne donnent pas de grosses différences dans les estimations et aucun critère ne s'impose, dans le cas des données traitées, permettant de choisir un modèle plutôt qu'un autre. Cela tient sans doute à la relativement faible fonction d'auto-corrélation sur les résidus. En revanche, même si la composante résiduelle reste assez faible, le fait de modéliser cette composante entraine des variations dans les estimations. L'introduction d'un terme $\Theta_6 B^6$ dans le polynôme ARMA améliore l'ajustement des courbes mais modifie aussi les estimations de certains effets, comme ceux dûs aux hausses tarifaires. Ces effets interviennent souvent avec six ou douze mois d'intervalle, ce qui en rend difficile une évaluation précise et tend à confondre périodicité et effet de perturbation.

En ce qui concerne la comparaison avec d'autres modèles d'analyse et de prévision, celle-ci ne pourrait être vraiment effectuée que par des études fondées sur de nombreuses simulations à partir de modèles d'évolution et de perturbations préétablis et qu'il s'agirait de réestimer au mieux.

Cette étude a été menée conjointement avec le Service d'Analyse Economique et du Plan du Ministère des Transports. Nous remercions MM. MALAMOUD, PERROT et TAROUX pour leurs conseils et leur aimable collaboration.

REFERENCES

(1) - R. ASTIER, Ch. DUHAMEL - Etude de séries chronologiques fréquemment perturbées (1982) CEMS - contrat avec le SAE.

(2) - BOX, TIAO - Intervention analysis with applications to economic and environmental problems (1975) JASA 70 n° 349

(3) - J.P. INDJEHAGOPIAN - Analyse des interventions sur les prévisions issues de modèles ARIMA (1980) - CERESSEC

(4) - BOX, JENKINS - Time Series analysis ; forecasting and control (1976) Holden-days

(5) - GRANGER, NEWBOLD - Forecasting economic time series - Academic Press (1980)

(6) - D.A. PIERCE - Least square estimation in the regression model with ARMA errors Biometrika (1971), 58

- Distribution of residual autocorrelations in the regression model with ARMA errors, JRSS-B, 33 p.140 (1971)
- Testing goodness of fit for the distribution of errors in regression models, BIOMETRIKA (1979), 66, 1
- Some recent developments in seasonal adjustment, in : Directions in time series (1978)