DÉCEMBRE 1995
ISBN 2-11-086017-0

Ministère
de l'Équipement,
du Logement,
des Transports
et du Tourisme

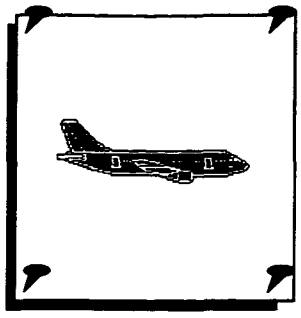
CONSTRUCTION D'UN OUTIL DE SUIVI ET DE PRÉVISION À COURT TERME DU TRAFIC AÉRIEN DOMESTIQUE

Stage de Benoît PINON (ENAC)
sous la direction de Ruth BERGEL (OEST)
en collaboration avec
Philippe CREBASSA (DGAC)

OEST - DGAC

OEST
10380

toire Économique et Statistique des Transports
2055 PARIS - La DEFENSE Cedex 04 Téléphone (1) 40 81 21 22 Télécopie (1) 40 81 17 71



UN OUTIL DE SUIVI ET DE PRÉVISION À COURT TERME DU TRAFIC AÉRIEN DOMESTIQUE

Ruth BERGEL,
*Philippe CREBASSA**
*et Benoit PINON***

Un modèle de suivi mensuel du trafic aérien domestique a été constitué à partir des statistiques aéroportuaires de trafics de passagers effectués sur les lignes intérieures françaises, toutes compagnies confondues.

Cet outil, qui apparaît robuste pour le trafic domestique hors radiales soumises à la concurrence, peut être utilisé pour une estimation mensuelle rapide à partir de cinq liaisons intérieures et comme indicateur avancé par une prévision à quelques mois.

Ce travail[1], qui a été réalisé dans le cadre d'une collaboration entre la DGAC (STA), l'ENAC et l'OEST, a permis de construire un modèle de suivi du trafic domestique. Il fournit une estimation rapide et robuste du trafic en le reliant à celui d'un petit nombre de liaisons intérieures ; il peut être utilisé comme indicateur avancé par une prévision à très court terme.

Son élaboration a comporté trois étapes : nettoyage de la base de données initiale pour en extraire une série estimative du trafic domestique total, construction d'un modèle qui relie le trafic total à celui d'un petit nombre de liaisons intérieures, enfin validation de cet outil pour le suivi et pour la prévision à court terme.

Une base de données volumineuse, allégée et redressée

La base de données initiale portait sur des données de trafics de passagers, détaillés par ligne, pour un ensemble de 3 000 lignes, par mois sur une période allant de janvier 1987 à décembre 1994. Un fichier allégé a été constitué par application de deux critères portant sur la stabilité des lignes (niveau minimum de trafic) et la fiabilité des données. Cette base de données allégée, puis redressée (l'échantillon des 60 lignes représentant 87 à 88% du trafic annuel sur la période), a servi à la constitution du modèle. Une approximation du trafic domestique total a été fournie dans le même temps, en rythme mensuel sur les 8 ans.

Huit radiales importantes, déjà soumises à la concurrence aérienne intérieure (Orly-Nice en 1992, Orly-Toulouse, Orly-Marseille, Orly-Bordeaux, Orly-Montpellier et Orly-Strasbourg courant 1995) ou susceptibles de l'être dans un avenir proche (Orly-Hyères et Orly-Lyon) ont été dans un premier temps exclues du champ de la modélisation.

Hormis huit radiales particulières...

Il s'est donc agi de relier le trafic domestique total *moins* les huit lignes citées ci-dessus à un petit nombre de lignes issues des 52 lignes restantes.
Une classification ascendante hiérarchique a été effectuée, pour tenter de dégager des groupes de lignes homogènes. Or, même après correction de quelques valeurs aberrantes, les lignes se trouvent regroupées en fonction de leurs données atypiques (évènements locaux exceptionnels, ou généraux comme la crise du Golfe fin 1990 début 1991), de sorte que cette tentative de classification a été abandonnée faute de temps suffisant pour corriger l'ensemble des valeurs atypiques.

* Direction générale de l'aviation civile (DGAC) - Service du transport aérien (STA).
** Ecole nationale d'aviation civile (ENAC).
1 Rapport d'étude disponible à l'OEST.

... le trafic est fonction de cinq liaisons...

Une régression avec stepwise du trafic domestique hors les huit radiales sur les 52 lignes restantes de la base allégée a permis de sélectionner cinq lignes qui suffisent à déterminer l'ensemble par relation statistique : Bastia-Nice, Biarritz-Orly, Bordeaux-CDG, Lyon-Toulouse, Orly-Rennes. Le modèle retenu est le suivant :

Trafic total hors radiales soumises à la concurrence = 0,021 + 0,092 Bastia-Nice*
(0,003) (0,023)
+ 0,305 Biarritz-Orly + 0,061 Bordeaux-CDG + 0,174 Lyon-Toulouse +0,136 Orly-Rennes.
(0,023) (0,010) (0,025) (0,010)

R2 = 0,920 DW = 1,841

** Les données sont exprimées en variation du logarithme.*

... et est mieux prévu en 1994 qu'avec une prévision directe

Ce modèle a ensuite été utilisé pour la prévision à court terme, par application des méthodes de Box et Jenkins à chacune des cinq séries de trafic sélectionnées par le modèle de suivi, puis par combinaison linéaire des prévisions obtenues pour l'année 1994 (les modèles SARIMA*, d'abord estimés sur 1987-1994, ont été réestimés sur 1987-1993, pour fournir une prévision pour l'année 1994).

On a également retenu un modèle SARIMA pour le trafic total hors radiales soumises à la concurrence, et comparé aux réalisations les prévisions obtenues par ces deux méthodes, pour l'ensemble de l'année 1994. On constate que, pour chacun des mois de l'année, la prévision obtenue en utilisant le modèle de régression est plus proche de la réalisation que la prévision directe. Alors que le trafic se redresse en 1994, la prévision par régression conduit à une sous-estimation du trafic de 1,5% contre 2,2% pour la prévision directe.

Un modèle moins robuste pour les huit autres radiales

Enfin, pour l'ensemble des huit radiales restantes, un modèle SARIMA a également été constitué. Mais à la différence du modèle précédent, son intérêt se limite au suivi conjoncturel sur le passé récent ; de fait, la robustesse du modèle sur ces lignes sujettes à des modifications sensibles n'est pas assurée. A titre d'exemple, une prévision de ce trafic sur l'année 1994 sous-estime de 5% le trafic effectivement réalisé.

Une correction des valeurs atypiques, et une classification des lignes à effectuer

Cet outil de suivi du trafic domestique mensuel hors radiales soumises à la concurrence, qui apparaît robuste (les cinq lignes sélectionnées sur la période 1987-1994 le sont aussi sur 1987-1993), peut être utilisé pour une estimation mensuelle rapide à partir des cinq liaisons retenues, et comme indicateur avancé par une prévision à quelques mois, par combinaison linéaire des prévisions à quelques mois réalisées pour chacune des cinq liaisons.
Le modèle pourra être amélioré après correction systématique des valeurs atypiques de la base de données allégée et redressée qui a servi à la constitution du modèle ; ce travail de correction devrait également permettre d'améliorer les modèles de prévision retenus. En particulier, une classification préliminaire des lignes permettrait d'identifier un certain nombre de groupes de lignes dont l'évolution du trafic soit similaire, et de retenir une ligne représentative de chaque groupe. ■

* Modèles ARIMA saisonniers (cf méthodologie détaillée dans le rapport d'étude disponible à l'OEST).

Que soient ici remerciés Mme Ruth Bergel et M. Philippe Crébassa qui ont eu la gentillesse de m'accorder beaucoup de leur temps précieux, ainsi que toutes les personnes de la DGAC et de l'OEST qui m'ont apporté éclaircissements et assistance.

TABLE DES MATIERES

INTRODUCTION

I - Objectifs

Cette étude vise à constituer un indicateur du **trafic aérien de passagers** en France, trafic domestique (**métropolitain**, plus exactement) et **commercial** (régulier et non régulier). Pour la DGAC, il s'agit à la fois d'effectuer un **suivi** de ce trafic, et d'en calculer des **prévisions à court terme**. L'OEST s'y intéresse également à des fins de suivi et de prévision, dans une optique multimodale notamment.

II - Comment obtenir des données de trafic aérien ?

Il existe deux sources :

- ✓ les compagnies aériennes
- ✓ les aéroports

(Remarquons que dans les deux cas le passager est comptabilisé car il est synonyme de revenu, par le biais de la taxe passager ou du billet acheté...)

En raison de l'intensification de la concurrence, le nombre de passagers transportés est devenu une information stratégique que les compagnies préfèrent ne pas divulguer.

Les aéroports tiennent une comptabilité statistique par ligne aérienne, grâce au formulaire rempli pour chaque vol et destiné à calculer la redevance passagers. Les trafics par aéroport et le trafic total qui nous intéresse sont déduits de ces statistiques par lignes, centralisées à la DGAC, qui constituent donc notre base de données.

III - Méthodologie

Un problème, qui justifie notre étude, se pose: tous les aéroports ne communiquent pas systématiquement leurs données de trafic, ou les fournissent avec quelques mois de retard, ce qui empêche d'actualiser le trafic total. Il faut donc bâtir un outil simple et rapide à mettre en oeuvre, qui fournisse une estimation du trafic total des derniers mois, et une prévision à quelques mois.

La méthode envisagée est d'établir une relation entre le trafic total et les trafics de quelques lignes (disons moins de 10) judicieusement choisies. L'intérêt est qu'il est alors facile d'obtenir les données récentes du petit nombre d'aéroports concernés, qui permettront d'en déduire les estimations et prévisions à court terme du trafic total. Le matériel requis devrait se limiter à un téléphone et une calculatrice. Et d'une bonne dose de vigilance et de bon sens, bien sûr, comme toujours quand il s'agit d'utiliser un modèle économétrique.

Dans un premier temps, il faut se pencher sur la base de données, la "nettoyer" , et déterminer ainsi, dans un premier temps (**partie I**) un ensemble de lignes susceptibles d'être sélectionnées pour le modèle.

On pourra ensuite choisir le modèle explicatif reliant le trafic total à quelques lignes représentatives (**partie II**) et construire un outil de prévision (**partie III**).

PARTIE I :

TRAITEMENT DE LA BASE DE DONNEES

Les services de l'administration sont très friands de statistiques - ils les collectent, les additionnent, les élèvent à la puissance n, en tirent la racine cubique et confectionnent de merveilleux diagrammes. Mais ce que vous ne devez jamais oublier c'est que chacun de ces chiffres est enregistré en première instance, par un garde-champêtre ou un autre bougre, qui n'inscrit que ce qu'il lui plaît d'inscrire.

SIR JOSIAH STAMP[1]

---

[1] cité in *Statistique* de T.H. Wonnacott et R.J. Wonnacott (Economica 91)

I - Description de la base de données

I-1- Le trafic par lignes aériennes

Nous disposons du trafic des lignes domestiques françaises en nombre de passagers par mois sur la période allant de janvier 1987 à décembre 1994. (Ce qui représente donc 96 mois). Rappelons qu'il s'agit du trafic commercial régulier et non régulier (charters).

Pour une liaison aérienne entre deux aéroports, sont supposées figurer dans la base les données de trafic départ + arrivée fournies par chaque aéroport.

Notons tout de suite une ambiguïté concernant le trafic envisagé pour une ligne. Quand on parle du trafic d'une ligne aérienne domestique on souhaite exclure les passagers en correspondance avec l'avion ou le train pour une autre destination. Or la détection des passagers en transit ne paraît envisageable que pour les correspondances avec un billet unique, ce qui ne concernait jusqu'à présent que les passagers en correspondance avec la même compagnie aérienne (La SNCF et Air Inter viennent de conclure un accord pour un billet unique en cas de correspondance avec le T.G.V. Nord à Charles De Gaulle). Les notions de trafic domestique et de transit sont donc plus floues qu'il n'y paraît. Mais cette difficulté à savoir quel trafic exact représentent les données a toutes les chances de persister, et ne gênent donc finalement pas trop la modélisation.

Un premier examen de la base de données révèle trois difficultés:

- ✓ On ne dispose pas toujours des deux sources.
- ✓ Quand elles existent, les deux sources donnent presque toujours des valeurs différentes (ce qui, pour la petite histoire, m'a longtemps fait croire à une distinction entre départs et arrivées, alors qu'en fait chacune des deux sources est sensée donner le trafic arrivée + départ...), et parfois incohérentes entre elles.
- ✓ La base de données présente de nombreux "trous"; en fait les mois manquants (mois où l'on ne possède pas la donnée) et les mois à trafic nul sont tous représentés par un "blanc".

Une étape de nettoyage de la base de données s'impose donc.

I-2- Le trafic total

Evidemment, les considérations concernant la signification précise des données restent valables.

Le trafic domestique mensuel se déduit par sommation du trafic par lignes. Or comme nous l'avons vu, la base de données est assez incomplète et une simple sommation est impossible.

Les seules données sur lesquelles nous pouvons nous appuyer sont des données annuelles fournies par la DGAC, que nous considérons comme fiables par manque d'information sur leur constitution.

II - Préparation de la base de données

Il s'agit de préparer la base de données de manière à ce qu'elle soit utilisable pour les études statistiques, c'est à dire qu'il ne faut retenir que les lignes susceptibles a priori (c'est à dire sans considérations statistiques de représentativité) d'être sélectionnées dans le modèle.

Ce filtrage des données de la base s'orientera autour de trois axes:

- ✓ stabilité du trafic de la ligne
- ✓ problèmes de données manquantes
- ✓ fiabilité des sources

II-1- Stabilité de la ligne

Il est certainement intéressant d'éliminer les lignes pour lesquelles le poids de l'aléatoire est trop fort, c'est-à-dire les lignes à faible trafic. L'évolution du transport aérien ou de la situation économique ont moins d'effet sur les variations de trafic de telles lignes que des phénomènes ponctuels que l'on peut qualifier d'aléatoires. (Ces considérations ne sont pas sans lien avec la loi des grands nombres: pour déterminer si un dé est ou non pipé, il faut procéder à un grand nombre de jets. Comme l'exprime Albert Jacquard: " Ainsi, paradoxalement, l'accumulation d'événements "au hasard" aboutit [...] à une répartition parfaitement prévisible des divers résultats possibles. Le "hasard" n'est capricieux que coup par coup; à long terme, ses interventions répétées créent un certain ordre, ou du moins un désordre suffisamment organisé [...].")

Nous avons choisi dans un premier temps un seuil bas de 1000 passagers par mois. Plus précisément nous avons sélectionné dans la base de données les lignes pour lesquelles une des deux sources au moins possède une moyenne de trafic mensuel (moyenne calculée sur 96 mois moins les mois manquants) supérieure ou égale à 1000 passagers. Ne sont retenues alors que 127 lignes (dont 6 ne possèdent qu'une seule source sur les 8 ans.)

II-2- Données manquantes

Pour des raisons de simplicité, nous avons ensuite sélectionné les lignes pour lesquelles l'une au moins des deux sources est complète (c'est-à-dire qu'aucun mois ne manque ou n'a un trafic nul; remarquons qu'après la sélection effectuée à l'étape précédente, on est à peu près assuré qu'un mois non renseigné correspond à une donnée manquante et non à un trafic nul).

On peut se consoler de l'élimination éventuelle de lignes importantes (c'est le cas par exemple de la ligne BÂLE-MULHOUSE PARIS ORLY), en considérant que les lignes éliminées sont celles pour lesquelles l'obtention des données pose des problèmes; gardons à l'esprit le but pratique de notre étude.

II-3- Fiabilité des sources

La nécessité de se pencher sur la fiabilité des sources provient de la constatation que les deux sources d'une ligne, pour un même mois, donnaient parfois des données très différentes.

Exemple de données incohérentes:
la ligne BERGERAC ORLY en novembre 1990

source	novembre 1990
BERGERAC	2245 pax
ORLY	837 pax

On peut envisager trois sources d'erreurs au moins:

✓ une erreur de saisie

Peut se produire à n'importe quel maillon de la chaîne entre le remplissage du formulaire de vol et l'informatisation des données.

✓ un décalage dans la période prise en compte

Si un aéroport fournit le trafic du 1er au 30 du mois et l'autre du 28 du mois précédent au 27 du mois en court, un décalage de trois jours se crée. Si le trafic de la ligne n'est pas très élevé et si, par exemple le nombre de week-ends n'est pas le même au cours des deux périodes, ce décalage peut induire une différence de trafic notable.

✓ une mauvaise prise en compte du transit

Dans le cas où l'un des deux aéroports est une plate-forme de correspondance, les données de l'autre aéroport peuvent être "gonflées" si il affecte tous les passagers au départ à la ligne, sans tenir compte de leurs correspondances. Cette explication devrait se traduire par un écart de signe presque constant entre les deux sources; dans ce cas on pourrait considérer que la source basse est la plus fiable.

En dehors de ce dernier cas il semble très difficile d'établir si une source est plus fiable que l'autre. Il faudrait étudier chaque cas d'espèce, se renseigner auprès des aéroports concernés... Ce qui serait trop fastidieux dans le cadre de notre étude. Après réflexion nous avons renoncé aussi à privilégier certains aéroports systématiquement. (Ce que fait la DGAC pour les aéroports Parisiens, par nécessité de cohérence interne dans sa gestion statistique).

Nous avons établi pour chaque ligne un indicateur de la cohérence de ses deux sources: la moyenne, calculée pour chaque mois où les deux sources sont renseignées, d'un écart relatif défini par:

$$ER = \frac{|source1 - source2|}{moyenne(source1, source2)}$$

N'ont alors été retenues que les lignes possédant leurs deux sources, et dont la moyenne de l'écart relatif est inférieure à 5%. Elles sont au nombre de 65,

possédant chacune deux sources, dont l'une au moins est complète. Bien sûr, cela ne donne pas une totale assurance de cohérence, dans la mesure où il s'agit d'une moyenne (calculée de surcroît sur un grand nombre de mois, parfois 96), mais les forts écarts ponctuels apparaîtront dans l'étude statistique comme des points aberrants susceptibles d'être corrigés.

Aucune de ces 65 lignes ne possède une source inférieure à l'autre pour plus de 90% des mois où elles sont toutes deux renseignées. Le cas de figure évoqué plus haut ne semble donc pas se présenter.

Faute de pouvoir décider qu'une source est plus fiable qu'une autre, nous avons choisi de prendre comme série de trafic pour une ligne aérienne *la moyenne de ses deux sources quand celles-ci sont complètes* (c'est-à-dire renseignées pour les 96 mois), ou *l'unique source complète le cas échéant.*

A partir de la base obtenue par application de ce principe qui est un pis-aller, nous avons enfin décidé d'augmenter le seuil bas de trafic à 2000 passagers, ce qui amène le nombre de lignes à 60. Il faut en effet réaliser un compromis entre la richesse de la base et son nombre de degrés de liberté (nombre de mois - nombre de lignes) qui doit être suffisant pour rendre les résultats statistiques significatifs.

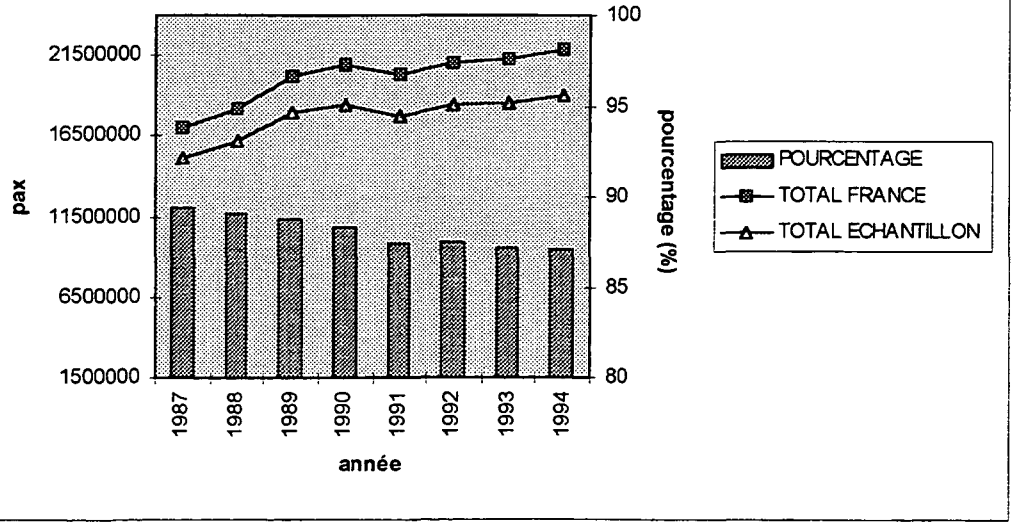
Se pose maintenant le problème de la série de trafic total (France entière).

II-4- La série de trafic total

La partie utilisable de la base de données se réduit donc à 60 lignes sur les quelques 3000 lignes de départ. On ne peut donc pas obtenir directement la série de trafic total par sommation. Il faut savoir quelle part du trafic total représente notre échantillon afin d'opérer un redressement. Nous allons utiliser pour cela la série de trafic total annuelle dont on dispose.

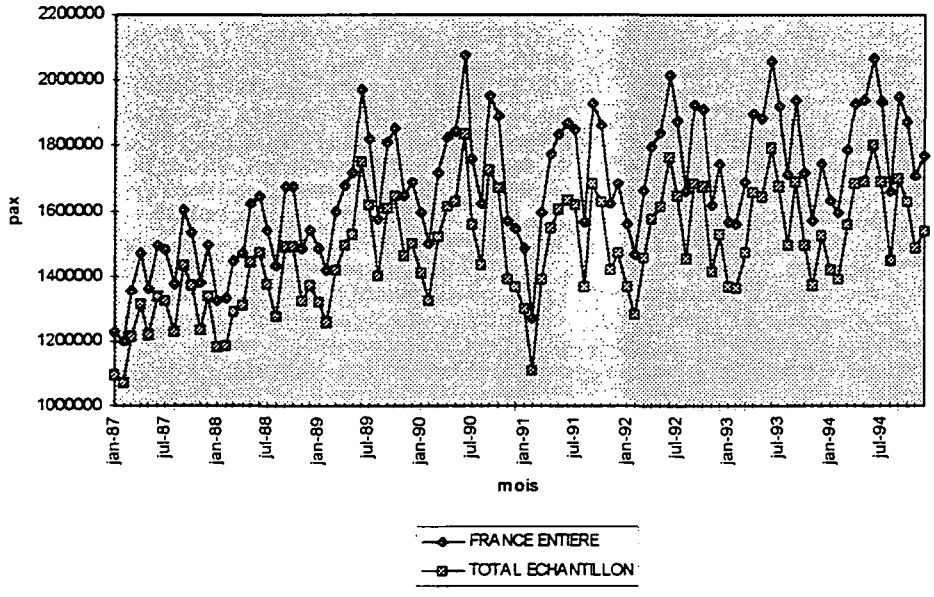
On calcule le trafic de notre échantillon pour chaque année, en sommant les trafics des 60 lignes mois par mois, puis en sommant les douze mois de l'année. On peut ainsi obtenir la part annuelle de l'échantillon dans le trafic total. (fig1)

fig1 : Part de l'échantillon dans le trafic France entière



On en déduit alors une approximation de la série de trafic mensuel France entière par redressement de la série mensuelle de l'échantillon, année par année, grâce aux pourcentages calculés. (fig2)

fig2 : Séries mensuelles



Résumons alors ce dont nous disposons pour l'étude statistique qui va suivre:

- ✓ Une base de données de trafic par lignes, en nombre de passagers mensuel, comprenant 60 lignes et couvrant la période de janvier 1987 à décembre 1994.
- ✓ Une approximation du trafic domestique en nombre de passagers mensuel, sur la même période.

Nous pouvons maintenant entamer la modélisation.

PARTIE II :

CONSTRUCTION D'UN OUTIL DE SUIVI

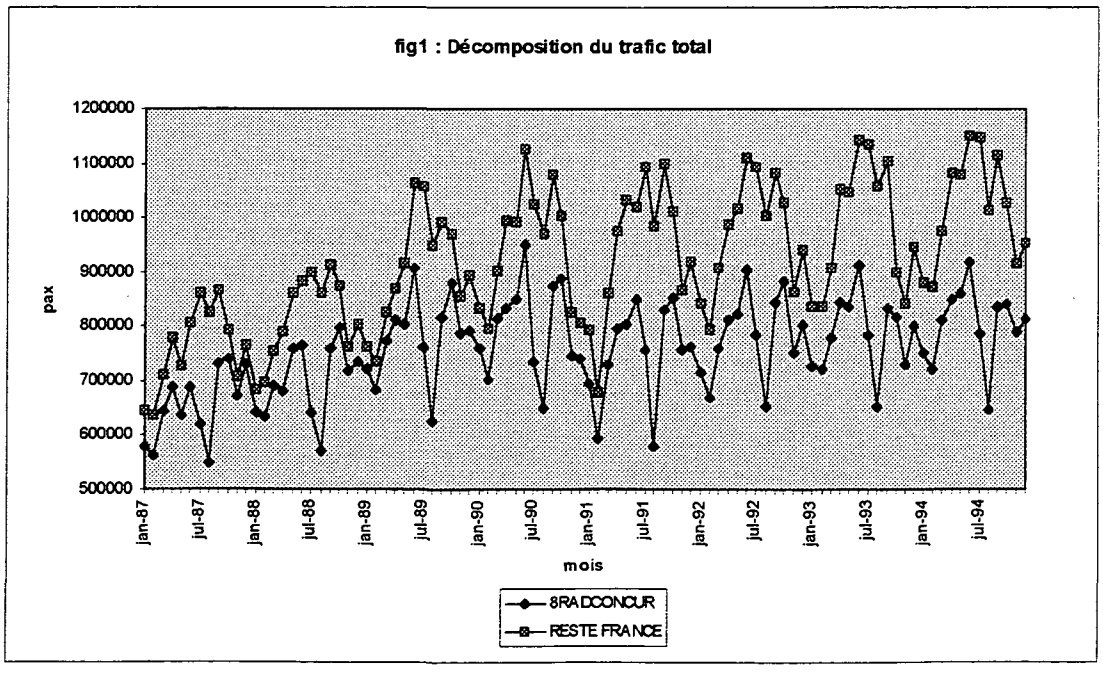
I- Choix du modèle

Rappelons l'objectif que l'on s'est fixé: trouver une relation entre le trafic domestique total et le trafic de quelques lignes, qui permettra de faire du suivi et de la prévision à court terme. Pour que cette relation, une fois obtenue, reste valable, il faut que la structure du transport aérien domestique français présente un minimum de stabilité. Or un phénomène nouveau est apparu il y a quelques années et a connu une récente intensification : la concurrence entre compagnies aériennes sur le marché domestique français.

Selon l'avis de spécialistes de la DGAC, ses effets ont été et resteront marginaux à Roissy Charles de Gaulle, au moins dans le moyen terme. En revanche les lignes ORLY NICE et plus récemment ORLY TOULOUSE BLAGNAC et ORLY MARSEILLE semblent très affectées par ce changement. Nous avons donc jugé qu'il était préférable d'étudier séparément certaines lignes radiales importantes soumises à la concurrence ou susceptibles de l'être dans un avenir proche. Il s'agit de :

1. ORLY BORDEAUX MERIGNAC
2. ORLY HYERES LE PALYVESTRE
3. ORLY LYON SATOLAS
4. ORLY MARSEILLE PROVENCE
5. ORLY MONTPELLIER FREJORGUES
6. ORLY NICE COTE D'AZUR
7. ORLY STRASBOURG ENTZHEIM
8. ORLY TOULOUSE BLAGNAC

Nous sommes donc amenés à réviser légèrement notre objectif : il s'agit de relier le trafic domestique total *moins* les huit lignes citées ci-dessus (qui constituent un ensemble que nous appellerons dans la suite 8RADCONCUR) à un petit nombre de lignes issues des 52 lignes restantes.



I-1- Transformation des données

On constate nettement sur la figure 1, et c'est vrai aussi bien sûr pour les lignes, individuellement, que le trafic en nombre de passagers comporte une composante saisonnière de période 12 (un an), comme c'est presque toujours le cas pour une série temporelle de base mensuelle. Notons que le mois d'août est un mois bien plus creux pour 8RADCONCUR que pour la somme des autres lignes (on peut suggérer que parmi ces dernières figurent bon nombre de lignes desservant des régions de tourisme estival). On conçoit que ces différences de saisonnalité existent aussi à l'intérieur du groupe des autres lignes. Si bien que selon le mois, le poids relatif des lignes varie. Il faudra donc tenir compte du mois de l'année dans la relation, ou ce qui est plus simple, travailler sur des données désaisonnalisées.

Un moyen simple de neutraliser la saisonnalité est de considérer des taux d'accroissement:

$T_t = \frac{Y_t - Y_{t-s}}{Y_{t-s}}$ où s est la saison.

Ici s=12, il s'agit donc d'une variation relative de trafic en % par rapport au même mois de l'année précédente.

En fait, de manière classique on utilise des données en *delta-log*, c'est-à-dire passer de Y_t à Z_t défini par :

$Z_t = \ln(Y_t) - \ln(Y_{t-12})$

C'est une transformation assez proche du taux d'accroissement.

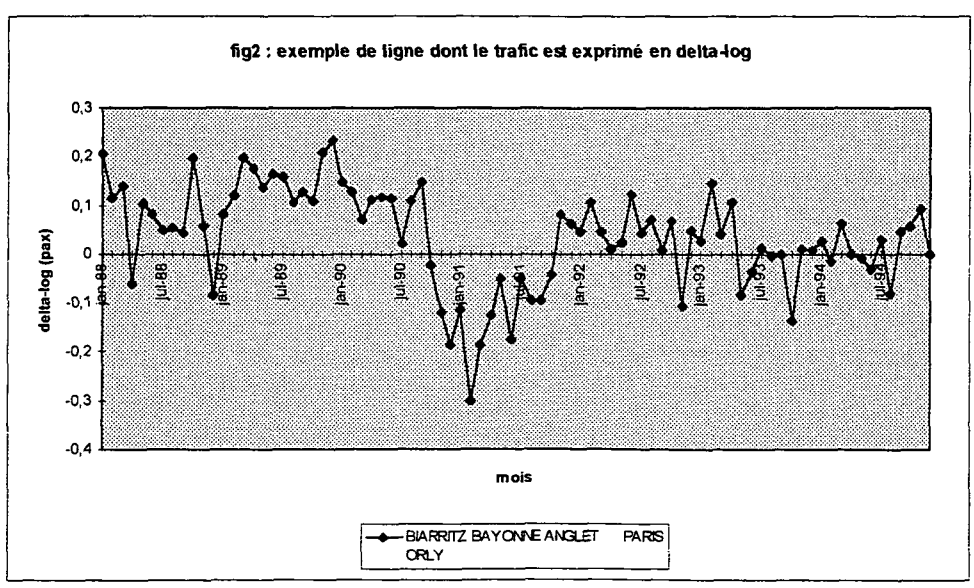
En effet :

$Z_t = \ln(Y_t) - \ln(Y_{t-12}) = \ln \frac{Y_t}{Y_{t-12}} = \ln \frac{Y_{t-12} + \Delta Y}{Y_{t-12}} = \ln(1 + \frac{\Delta Y}{Y_{t-12}})$

soit :

$Z_t \approx \frac{\Delta Y}{Y_{t-12}}$ si $\Delta Y << Y_{t-12}$, ce qui n'est pas toujours le cas néanmoins, il faut s'en méfier.

La période couverte par les données est donc réduite à 84 mois, de janvier 1988 à décembre 1994 (on ne peut pas, bien sûr, calculer de delta-log pour la première année.)



Nous travaillerons désormais avec cette nouvelle base de données, constituée de 52 lignes, dont le trafic mensuel est fourni en delta-log, sur la période allant de janvier 1988 à décembre 1994 (nous appellerons trafic "total" le trafic domestique France entière *moins* le trafic de 8RADCONCUR).

Avant de sélectionner des lignes explicatives, il convient de spécifier la classe du modèle que nous allons choisir.

I-2- Spécification du modèle

Nous chercherons une relation entre le trafic total et le trafic de lignes de la base (tous exprimés en delta-log) sous une forme linéaire.

C'est à dire que nous utiliserons le modèle de régression multiple classique :

$$Y_t = a_0 + a_1 X_{1t} + \ldots + a_k X_{kt} + \varepsilon_t \qquad \text{pour t=1,...,n}$$

avec:

Y_t : trafic total au mois t

X_{it} : trafic de la ligne i au mois t, avec $i \in [1,k]$. (On vérifie que les lignes X_i sont linéairement indépendantes dans $\mathbf{R}^n$).

k : nombre de lignes sélectionnées dans le modèle (k~5)

$a_0, a_1, \ldots, a_k$: paramètres du modèle, fixes mais inconnus

ε_t : terme aléatoire suivant une loi normale $N(0,\sigma)$. On suppose de plus que $\forall t \neq t'$ $\text{cov}(\varepsilon_t, \varepsilon_{t'}) = 0$.

Le modèle peut s'écrire matriciellement :
$Y = Xa + \varepsilon$
Les différentes hypothèses se traduisent par :

- ✓ X'X est inversible
- ✓ $\varepsilon / X, a, \sigma^2 \sim N(0, \sigma^2 I_n)$ (contient les hypothèses d'homoscédasticité et de non-corrélation des erreurs)

Ces hypothèses permettent l'application de la méthode des moindres carrés et la connaissance de la distribution des estimateurs.

Tout le problème est donc maintenant de choisir les lignes permettant d'obtenir la meilleure régression possible.

II - Choix des lignes

Plutôt que de chercher directement les lignes parmi les 52 candidates, il pourrait être intéressant de dégager des groupes de lignes homogènes.

II-1- Classification des lignes

On pourrait ainsi régresser le total de chaque groupe sur quelques lignes du groupe. L'intérêt est qu'on peut espérer, du fait de l'homogénéité interne du groupe, trouver facilement des lignes donnant une très bonne modélisation. Il suffirait ensuite de sommer les sous-modèles. Restent deux interrogations : de tels groupes existent-ils, et si c'est le cas, le résultat obtenu par sommation sera-t-il meilleur que celui du modèle découlant d'une recherche directe des lignes explicatives ?

Il serait préférable également de trouver *a posteriori* une caractéristique qualitative pour chaque groupe (c'est-à-dire de pouvoir interpréter économiquement la classification; déterminer, par exemple, que les lignes sont regroupées selon leur longueur, ou leur trafic en nombre de passagers, selon qu'elles sont radiales ou transversales, selon l'existence ou non de concurrence d'un autre mode de transport etc...), susceptible de fournir une règle de décision permettant d'affecter une ligne à un groupe; cela éviterait de relancer la procédure statistique de classification lors du recalibrage du modèle. Pouvoir étudier des groupes distincts de ligne est en plus intéressant sur le plan de la connaissance du transport aérien.

On a employé la méthode de classification ascendante hiérarchique à l'aide du critère de Ward[2].

Présentation de la méthode.

Le principe d'une classification ascendante est le suivant:

On associe aux données le nuage de points $N=\{X_1,...,X_p\}$ dans $\mathbf{R}^n$ (où X_i représente une ligne aérienne, p=52 est le nombre de lignes, n=84 est le nombre de mois).

L'algorithme de classification ascendante hiérarchique est itératif. A l'étape courante, on part d'une partition du nuage N des p lignes en k classes $G_1,...,G_k$ et on regroupe les deux classes les plus proches au sens d'une distance D. A l'étape initiale chaque ligne forme une classe, et à l'étape finale il n'y a plus qu'une seule classe. Restent à définir la distance D et le critère qui permet de décider du nombre de classes optimum.

Le centre de gravité du nuage N est le point g défini par :

$$g = \frac{1}{p}\sum_{i=1}^{p} X_i.$$

C'est en quelque sorte une ligne moyenne. La dispersion du nuage de points autour de son centre de gravité est mesurée à l'aide de l'*inertie* totale du nuage N, définie par :

$$I(N,g) = \frac{1}{p}\sum_{i=1}^{p} d^2(X_i,g)$$

où d est la distance euclidienne dans $\mathbf{R}^n$.

Considérons une partition du nuage N des p lignes en k classes $G_1,...,G_k$ d'effectifs respectifs $p_1,...,p_k$ et de centres de gravité $g_1,...,g_k$.

L'inertie totale du nuage N se décompose de la manière suivante :

$$I(N,g) = \sum_{i=1}^{k}\left(\frac{p_i}{p}\right)d^2(g_i,g) + \sum_{i=1}^{k}\left(\frac{p_i}{p}\right)I(G_i,g_i)$$

Le premier terme de droite s'appelle l'inertie inter-classes et mesure la manière dont les classes diffèrent les unes des autres. Ce terme est noté $I(G_1,...,G_k)$ et représente l'inertie expliquée par la partition. Le second terme de droite s'appelle l'inertie intra-classes et mesure l'homogénéité des classes.

On mesure la qualité d'une partition à l'aide du rapport inertie inter-classes sur inertie totale.

Lorsque dans la partition $G_1,...,G_k$ on regroupe deux classes G_i et G_j, il y a diminution de l'inertie inter-classes. Cette diminution :

$$D(G_i,G_j) = I(G_1,...,G_i,...,G_j,...,G_k) - I(G_1,...,G_i \cup G_j,...,G_k)$$

se calcule et vaut :

$$D(G_i,G_j) = \frac{p_i p_j}{p(p_i+p_j)}d^2(g_i,g_j)$$

Ce critère, appelé critère d'agrégation de Ward, mesure la distance entre les classes G_i et G_j.

A chaque étape de l'algorithme, la distance minimum (soit le critère de Ward minimum) entre deux classes augmente. On choisit la partition qui correspond à une augmentation brutale du critère, soit à une perte importante d'inertie inter-classes (soit encore à une diminution importante de l'hétérogénéité entre les classes).

---

2 La présentation de la méthode est inspirée de *Méthodes statistiques en Gestion* (chap. 6) de Michel Tenenhaus

Voilà un extrait du résultat :

Classe	Aîné	Benjamin	Effectif	Critère de Ward
51	BIARRITZ-ORLY	PERPIGNAN-ORLY	2	0,00088
...				
8	Classe 22	Classe 11	31	0,03239
7	Classe 8	NANTES-CDG	32	0,03367
6	Classe 13	Classe 10	11	0,04059
5	Classe 7	Classe 6	43	0,05268
4	Classe 20	LILLE-MARSEILLE	3	0,05689
3	Classe 9	Classe 5	48	0,10484
2	Classe 4	BASTIA-NICE	4	0,10784
1	Classe 3	Classe 2	52	0,17244

On constate des discontinuités du critère de Ward au passage de 7 à 6 classes, de 4 à 3 classes et de 2 à 1 classe(s). On peut donc envisager des partitions en 2, 4 ou 7 classes.

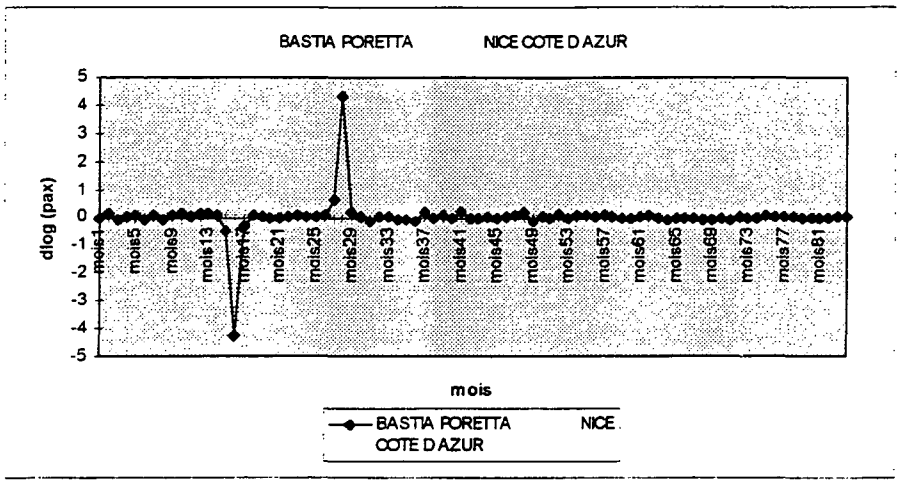
Voilà la répartition en 4 classes :

Classe 1	BASTIA - NICE
Classe 2	LILLE - MARSEILLE BASTIA - ORLY BASTIA - MARSEILLE
Classe 3	LYON - STRASBOURG NICE - STRASBOURG CHAMBERY - ORLY CDG - STRASBOURG LILLE - STRASBOURG
Classe 4	Le reste des lignes

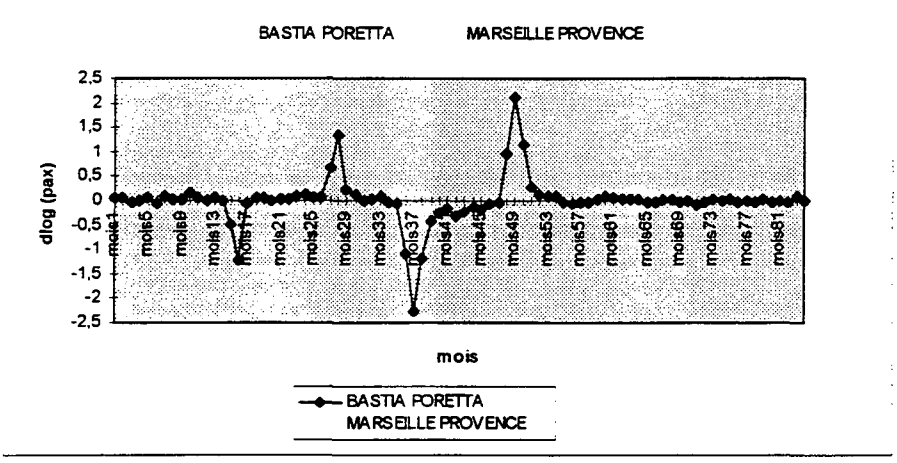
Il s'agit maintenant d'interpréter cette classification. Pour cela on peut commencer par visualiser quelques lignes. (voir page suivante)

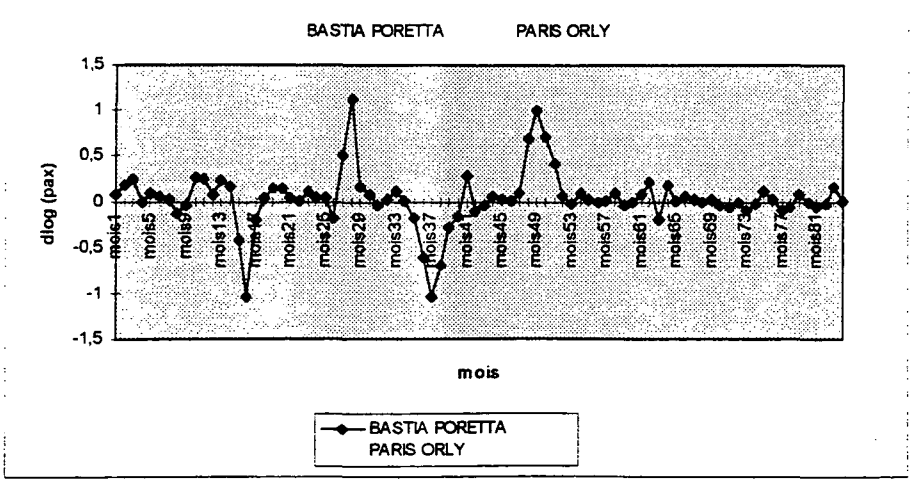
Fig 3 : Quelques lignes ressortant de la classification

Classe 1 :



Classe 2 :





On constate (et les autres graphiques le confirmeraient) que les lignes sont regroupées en fonction de leurs données atypiques. Des événements exceptionnels, locaux comme à Bastia en mars et avril 1989, ou généraux comme la crise du Golfe autour de janvier 1991, causent des valeurs anormales qui ont un poids très fort dans les calculs de distance de l'algorithme de classification.

Il faudrait donc supprimer les mois atypiques ou en corriger les valeurs puis relancer une classification. Le manque de temps nous a contraint d'y renoncer mais c'est une voie intéressante.

Nous nous sommes donc contentés d'une sélection directe à partir de l'ensemble des 52 lignes candidates.

II-2- Sélection des lignes[3]

Les méthodes de sélection automatique de variables explicatives les plus utilisées sont la régression pas à pas ascendante (*forward stepwise selection*), la régression pas à pas descendante (*backward stepwise selection*), ou une combinaison des deux.

- ✓ **La sélection pas à pas ascendante** est une méthode itérative consistant à sélectionner à chaque étape la variable explicative maximisant le R^2 de Y (la variable dépendante) avec toutes les variables sélectionnées aux étapes précédentes et la nouvelle variable choisie, tant que l'apport marginal de cette dernière est significatif. Si au cours d'une étape une des variables déjà sélectionnées devient non significative, elle est alors éliminée.

Coefficient de détermination R^2

Rappelons que :

$$Y_t = a_0 + a_1 X_{1t} + ... + a_k X_{kt} + \varepsilon_t .$$

On note alors :

$$\hat{Y_t} = \hat{a_0} + \hat{a_1} X_{1t} + ... + \hat{a_k} X_{kt}$$

avec : $\hat{a_0}, \hat{a_1}, ..., \hat{a_k}$ les estimations des paramètres $a_0, a_1, ..., a_k$ calculées à partir de l'échantillon

$\hat{Y_t}$ la valeur de Y pour l'observation t, calculée à l'aide d'une estimation du modèle

et : $e_t = Y_t - \hat{Y_t}$ le résidu de reconstitution de l'observation t

On peut montrer la formule de décomposition :

$$\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2 = \sum_{t=1}^{n} \left(\hat{Y_t} - \bar{Y}\right)^2 + \sum_{t=1}^{n} e_t^2$$ où $\bar{Y}$ désigne la moyenne des Y_t.

Le coefficient de détermination R^2 est défini par :

$$R^2 = \frac{\sum_{t=1}^{n} \left(\hat{Y_t} - \bar{Y}\right)^2}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2} = 1 - \frac{\sum_{t=1}^{n} e_t^2}{\sum_{t=1}^{n} \left(Y_t - \bar{Y}\right)^2}$$

Il mesure la part de variation totale expliquée par les variables $X_1, ..., X_k$.

On rencontre aussi le R^2 *ajusté*. Il s'agit de tenir compte du nombre d'observations n et du nombre de variables explicatives k :

$$R^2 \text{ ajusté} = 1 - \frac{\sum e_t^2 / (n-k-1)}{\sum \left(Y_t - \bar{Y}\right)^2 / (n-1)}$$

Grâce au R^2 ajusté, on peut comparer les pouvoirs explicatifs de différents modèles.

Les tests marginaux

Sur un modèle à k variables $Y = a_0 + a_1 X_1 + ... + a_k X_k + \varepsilon$, on peut étudier pour chaque variable X_i le test :

$$H_0 : a_i = 0$$
$$H_1 : a_i \neq 0$$

---

[3] Les rappels statistiques de tout ce paragraphe sont inspirés de *Méthodes statistiques en Gestion* de Michel Tenenhaus

Il s'agit de vérifier pour la variable X_i si, lorsqu'on passe du modèle à k variables explicatives au modèle simplifié obtenu en supprimant la variable X_i, il y a diminution significative de la qualité du modèle.
On utilise la statistique :

$t_i = \frac{\hat{a_i}}{s_i}$ où s_i est l'estimation de la variance de $\hat{a_i}$.

Sous l'hypothèse que H_0 est vraie, la statistique t_i suit une loi de Student à n-k-1 degrés de liberté. On rejette H_0 au risque de première espèce α si :

$|t_i| \geq t_{1-(\alpha/2)}(n-k-1)$ valeur lue dans la table de Student

ou de manière équivalente, si le niveau de signification du test :

$P=\text{Prob}(|T(n-k-1)| \geq |t_i|)$

est inférieur à α .

La statistique t_i mesure en fait *l'apport marginal* de la variable X_i à l'explication de Y. On peut en effet montrer que :

$$F_i = t_i^2 = \frac{S(X_1,...,X_k)-S(X_1,...,X_{i-1},X_{i+1},...,X_k)}{\hat{\sigma}^2}$$

où : $S(X_1,...,X_k)$ est la somme des carrés $\sum_{t=1}^{n}\left(\hat{Y_t}-\bar{Y}\right)^2$ expliquée par les variables $X_1,...,X_k$.

$\hat{\sigma}^2 = \sum_{t=1}^{n} e_t^2/(n-k-1)$ est l'estimation de la variance de l'erreur ε.

La statistique F_i s'appelle le F partiel.
Le critère utilisé pour sélectionner les variables revient à choisir à chaque étape la variable qui a le plus fort F partiel (ou $|t|$). En effet à l'étape courante (l variables déjà sélectionnées), on a :

$$F_i = t_i^2 = \frac{S(X_1,...,X_l,X_i)-S(X_1,...,X_l)}{\left(\sum\left(Y_t-\bar{Y}\right)^2-S(X_1,...,X_l,X_i)\right)/(n-l-2)}$$

$$= \frac{(n-l-2)[R^2(Y;X_1,...,X_l,X_i)-R^2(Y;X_1,...,X_l)]}{1-R^2(Y;X_1,...,X_l,X_i)}$$

La statistique F_i est donc une fonction croissante de $R^2(Y;X_1,...,X_l,X_i)$; les deux critères sont bien équivalents.

✓ Dans **la régression pas à pas descendante**, on part cette fois du modèle complet à 52 lignes, et on élimine à chaque étape la variable ayant le plus petit apport marginal (soit le plus petit F partiel), à condition qu'il soit non significatif. Si au cours d'une étape une des variables déjà éliminées devient significative, elle est alors réintégrée.

A l'étape courante (l variables encore dans le modèle) on a :

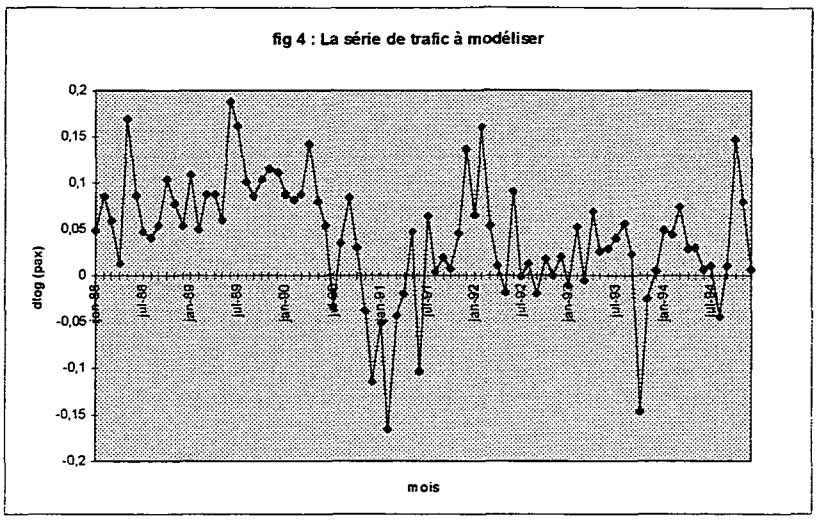
$$F_i = t_i^2 = \frac{R^2(Y;X_1,...,X_l)-R^2(Y;X_1,...,X_{i-1},X_{i+1},...,X_l)}{(1-R^2(Y;X_1,...,X_l))(n-l-1)}.$$

Cette méthode revient donc à supprimer à chaque étape la variable X_i telle que le R^2 diminue le moins possible.

Avant d'expérimenter ces méthodes sur notre base de données, on peut s'interroger sur l'effet d'éventuelles valeurs atypiques sur les résultats.

Sur le graphique (fig 4, page suivante) du trafic total (rappelons qu'il s'agit du trafic France entière *moins* 8RADCONCUR), semblent ressortir deux périodes atypiques caractérisées par une forte baisse de trafic: la crise du Golfe autour de janvier 1991, et les mouvements sociaux qui ont touché les aéroports parisiens en octobre et novembre 1993.

Ces valeurs anormales (c'est à dire éloignées des autres) vont "attirer" le plan des moindres carrés (qui ajuste au mieux les données) et vont donc avoir un fort poids sur les résultats de la régression, et sur les procédures de recherche des lignes optimisant la régression.



On peut considérer que la relation est faussée si, par exemple, toutes les lignes n'ont pas subi l'événement exceptionnel à l'origine de la valeur atypique. C'est le cas des mouvements sociaux de 1993 qui n'ont touché que les lignes radiales. Pour rendre compte de ce ces points atypiques, l'algorithme du stepwise risque de privilégier les radiales. La situation est un peut différente en ce qui concerne la guerre du Golfe, dont les effets sur le transport aérien domestique sont presque exclusivement dus à une crise économique, un ralentissement de la consommation, plutôt qu'aux menaces terroristes. Si l'on considère que toutes les régions françaises ont été affectées à peu près de la même façon par cette crise, on peut admettre pour simplifier que ses effets sur chaque ligne sont respectueux des caractéristiques de la ligne et donc à prendre en compte pour l'élaboration du modèle. (En outre, la correction systématique des effets de la guerre du Golfe sur toutes les lignes et le total serait extrêmement fastidieuse, quant à l'élimination pure et simple de cette période, elle réduirait le nombre de mois à 60 environ[4], ce qui réduirait le nombre de degrés de liberté du système complet à 52 lignes et rendrait plus délicate encore la sélection des lignes).

En conséquence, nous avons décidé de supprimer de la base de données les mois d'octobre et novembre des années 1993 et 1994[4].

Après plusieurs essais d'ajout et de retrait de lignes aux sélections obtenues par les méthodes de régression pas à pas, le meilleur modèle à 5 explicatives que l'on ait trouvé correspond aux lignes suivantes:

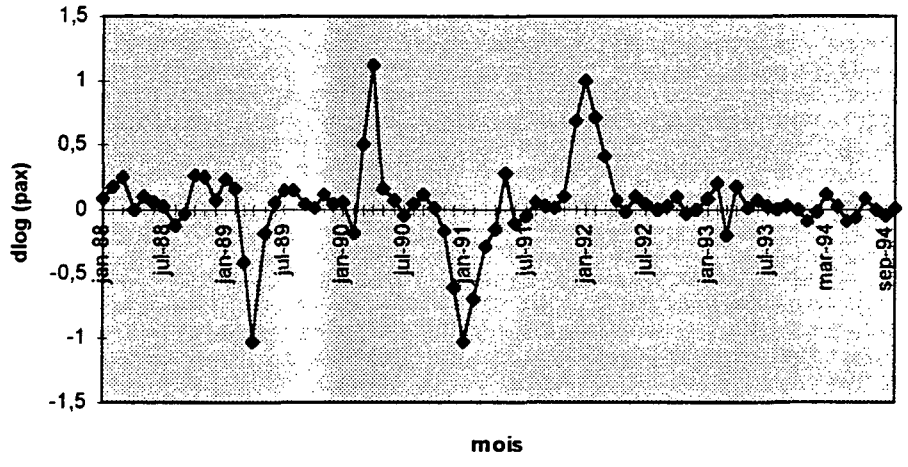
1. BASTIA ORLY
2. BIARRITZ ORLY
3. LYON NICE
4. LYON TOULOUSE
5. ORLY RENNES

---

[4] Quand un trafic en nombre de passagers est atypique pour un mois donné, il entraîne dans la base en delta-log une valeur aberrante pour le mois en question, et sa compensation au même mois de l'année suivante.

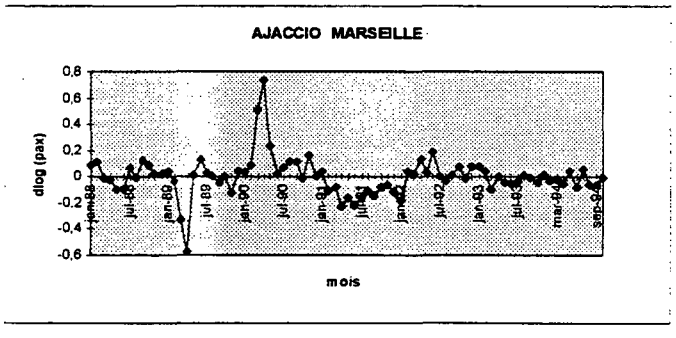
Sur le graphique de la série de trafic BASTIA ORLY (fig 5) on constate qu'outre la période de la crise du Golfe, les mois de mars et d'avril 1989 sont très atypiques.

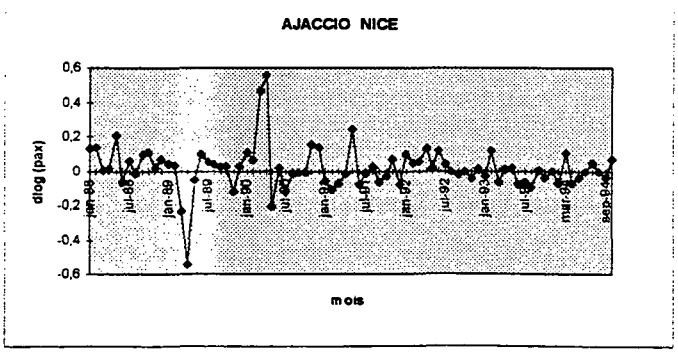
fig 5 : La ligne BASTIA ORLY

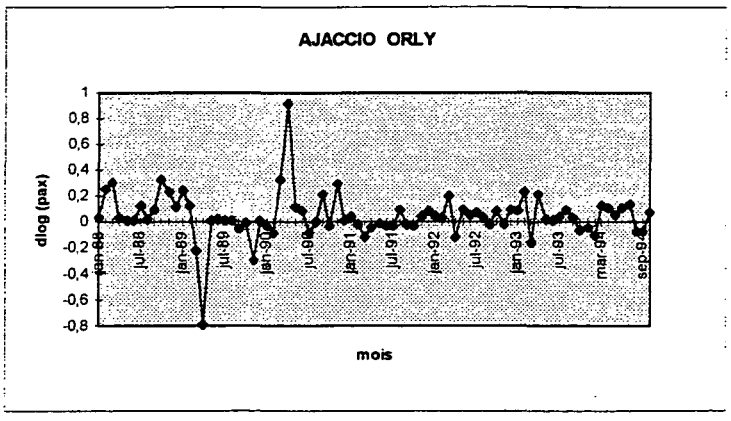


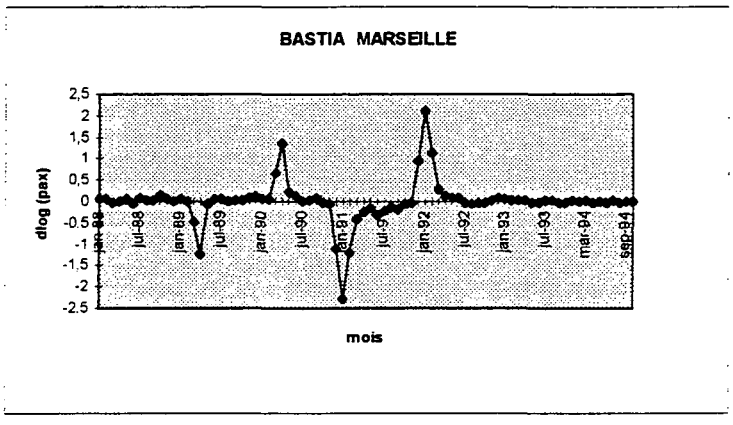
Les deux aéroports donnent pour ces mois là des données voisines; la chute de trafic de la ligne est donc à imputer à un événement exceptionnel. La figure 6 montre que c'est un événement qui a touché la Corse.

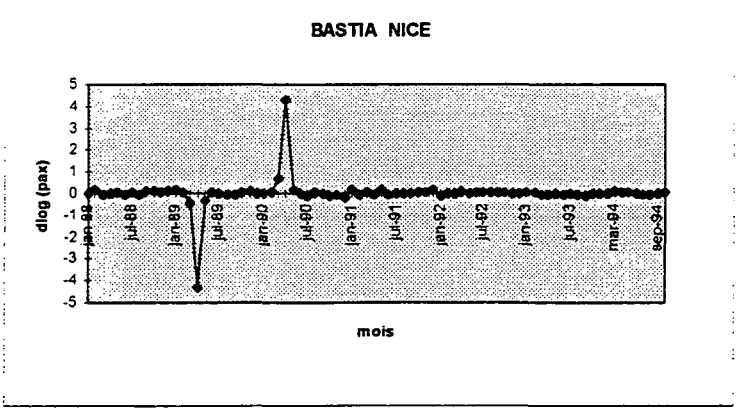
fig 6 : Les lignes d'AJACCIO et BASTIA









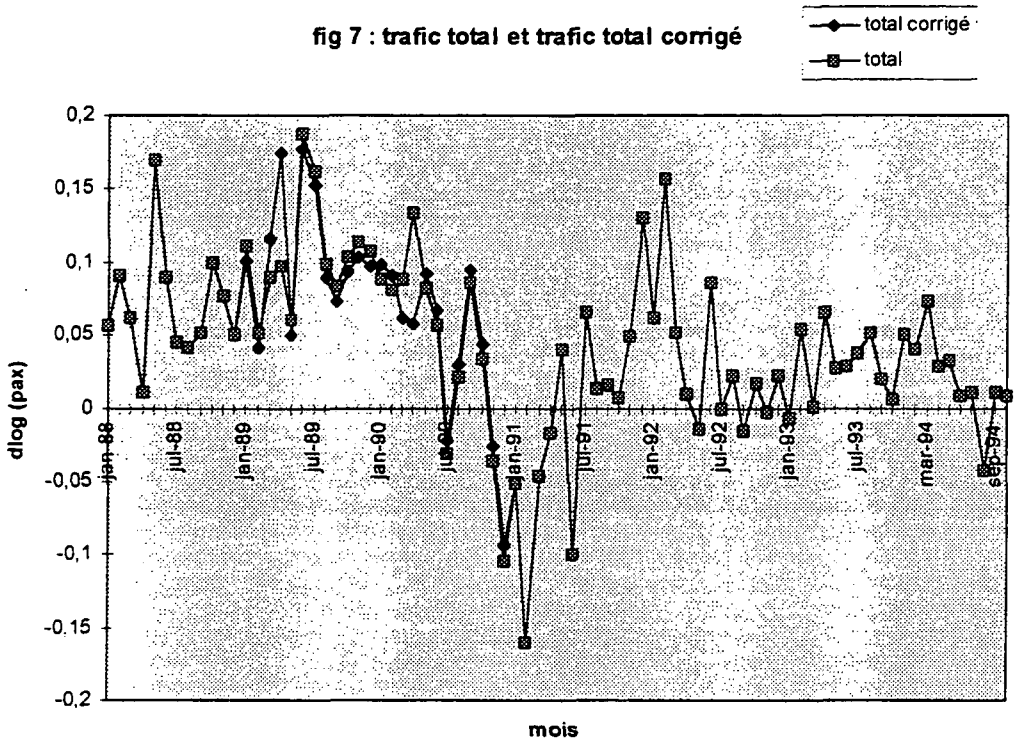


Or la moyenne de trafic de la somme de ces six lignes vers la Corse est environ de 110000 passagers par mois, ce qui représente un huitième du total à modéliser. L'impact de ces deux mois atypiques sur le trafic total est donc important, d'autant qu'ils influent sur l'ensemble des deux années 1989 et 1990 par le biais du pourcentage de redressement (voir partie I - II - 4) et du passage au delta-log (voir note p22).

Il serait donc intéressant de corriger les valeurs de trafic pour ces deux mois. Pour cela, nous avons choisi de remplacer, *dans la base de donnée en nombre de passagers*, les mois de mars et d'avril 1989 par la demi-somme des mêmes mois des années 1988 et 1990, pour chacune des six lignes évoquées ci-dessus. Il faut ensuite

reprendre les étapes décrites dans la partie I pour obtenir la nouvelle approximation de la série totale (la part de l'échantillon a changé pour l'année 1989), et pour la transformation en delta-log. La figure 7 confirme que la modification est de taille.

fig 7 : trafic total et trafic total corrigé



Il faut, avec la base de données corrigée (de laquelle on a aussi supprimé les mois d'octobre et novembre des années 1993 et 1994), relancer les procédures de sélection de lignes explicatives.

A partir des sélections fournies par les procédures de régression pas à pas, plusieurs tentatives d'améliorations nous ont conduits à retenir les 5 lignes suivantes:

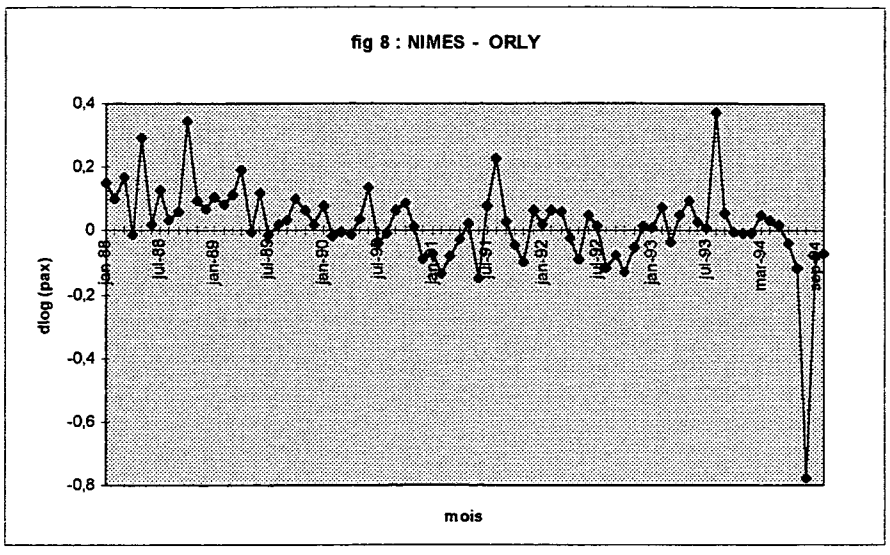
1. BIARRITZ ORLY
2. BORDEAUX PARIS CHARLES DE GAULLE
3. LYON TOULOUSE
4. NIMES ORLY
5. ORLY RENNES

Comme on pouvait s'y attendre, les lignes sélectionnées ont partiellement changé par rapport à notre première sélection.

La figure 8 montre que les mois d'août 1993 et 1994 sont fortement atypiques pour la ligne NIMES - ORLY. Afin de déterminer si ces mois ont joué un rôle important dans la régression, nous allons utiliser une mesure de l'influence d'une observation : le DFITS.

Pour le mois i, on définit :

$$DFITS_i = \frac{\hat{Y}_i - \hat{Y}_{(i)}}{\hat{\sigma}_{(i)}\sqrt{h_i}}$$



où :

$\hat{Y}(i)$ est la valeur calculée de Y pour le mois i sans utiliser ce mois pour le calcul des coefficients de régression.

$\hat{\sigma}(i)$ est l'estimation de l'écart type de ε, obtenue sans utiliser le mois i dans les calculs.

h_i est le **levier** (*leverage*) du mois i.

C'est une mesure de la distance entre le mois i $M_i=(X_{i1},...,X_{i52})$ et le mois moyen $\overline{M}=\left(\overline{X_1},...,\overline{X_{52}}\right)$:

$$h_i = \frac{1}{n}\left(1 + d^2(M_i, \overline{M})\right)$$

où $d(M_i, \overline{M})$ est la distance de Mahalanobis[5] entre les mois M_i et $\overline{M}$.

Une forte valeur du $|DFITS_i|$ indique donc des modifications importantes dans l'estimation des paramètres du modèle lorsqu'on retire le i-ème mois, et permet donc d'affirmer que ce mois est influent sur les résultats de la régression.

Le mois d'août 1994 possède le plus grand DFITS de tous les mois (il vaut 1.58740 le seuil, dans notre cas étant de l'ordre de 0.6). C'est donc bien un mois trés influent sur les paramètres de la régression, ce qui suggère de retirer les mois d'août 1993 et 1994.

On relance alors une fois de plus les procédures de sélection de lignes explicatives sur la nouvelle base de donnée comportant 78 mois (les mois d'août, octobre et novembre des années 1993 et 1994 ont été enlevés, et cette fois sont retenues:

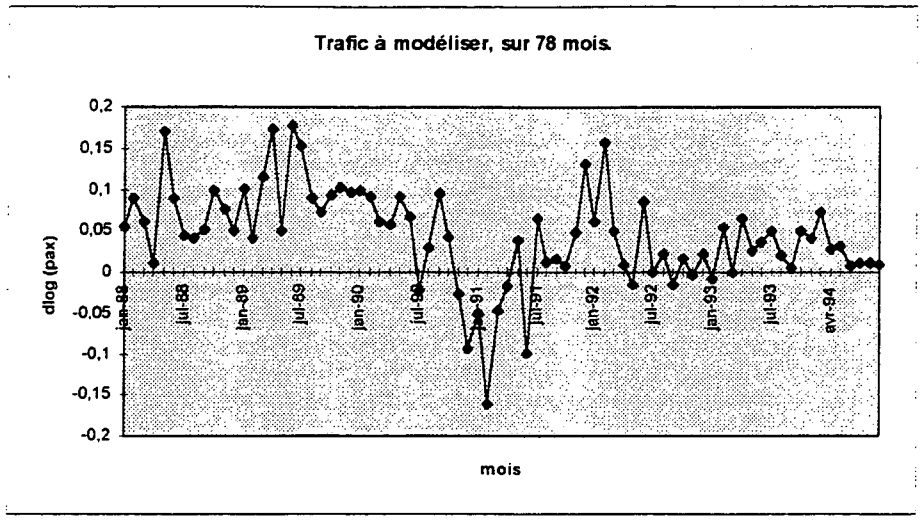
1. BASTIA PORETTA NICE COTE D'AZUR
2. BIARRITZ BAYONNE ANGLET PARIS ORLY
3. BORDEAUX MERIGNAC PARIS CHARLES DE GAULLE
4. LYON SATOLAS TOULOUSE BLAGNAC
5. PARIS ORLY RENNES ST JACQUES

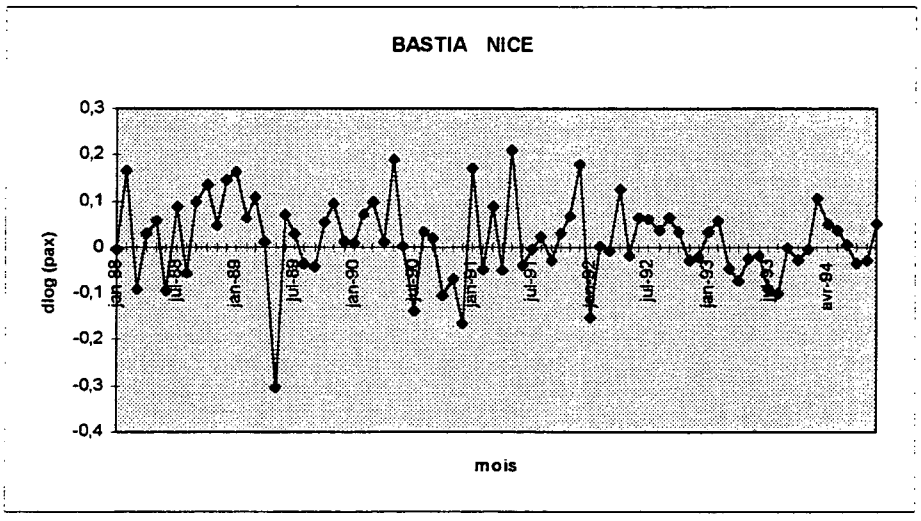
Les cinq séries sont représentées dans la figure 9.

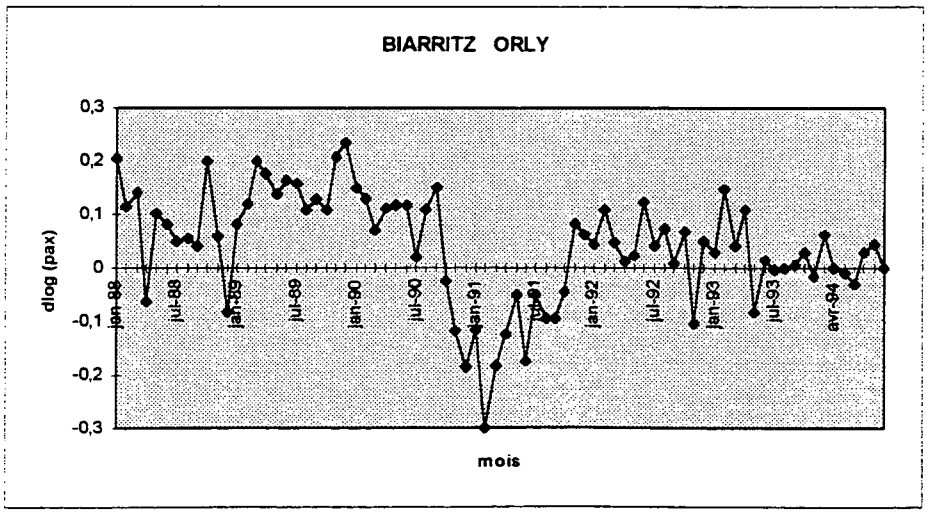
---

5 Voir *Méthodes statistiques en gestion* de Michel Tenenhaus, chap.6 p132

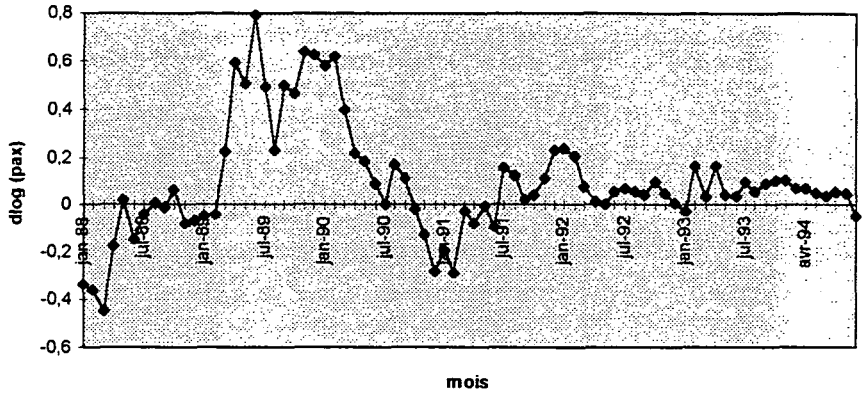
fig 9 : Le trafic à modéliser et les 5 lignes explicatives



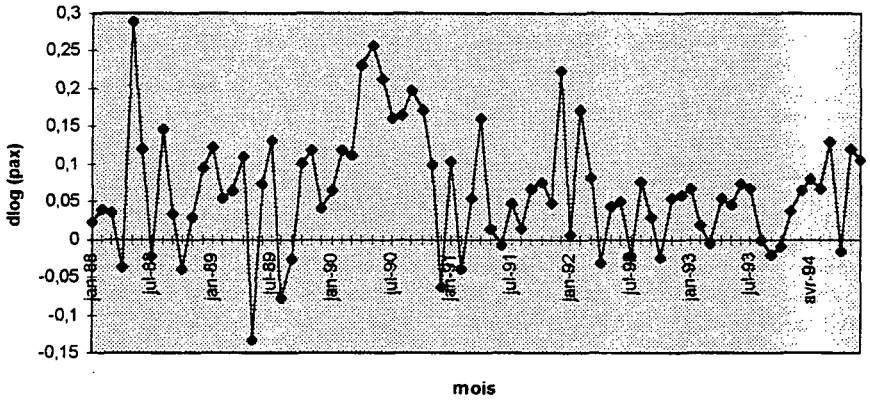




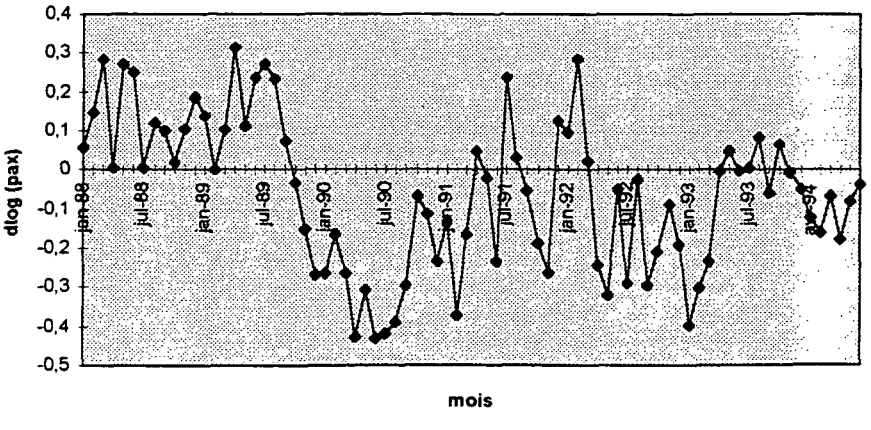
BORDEAUX CDG



LYON TOULOUSE



ORLY RENNES



III - Etude du modèle

III-1 Présentation des résultats

Nous présentons ci-dessous les résultats de la régression multiple :

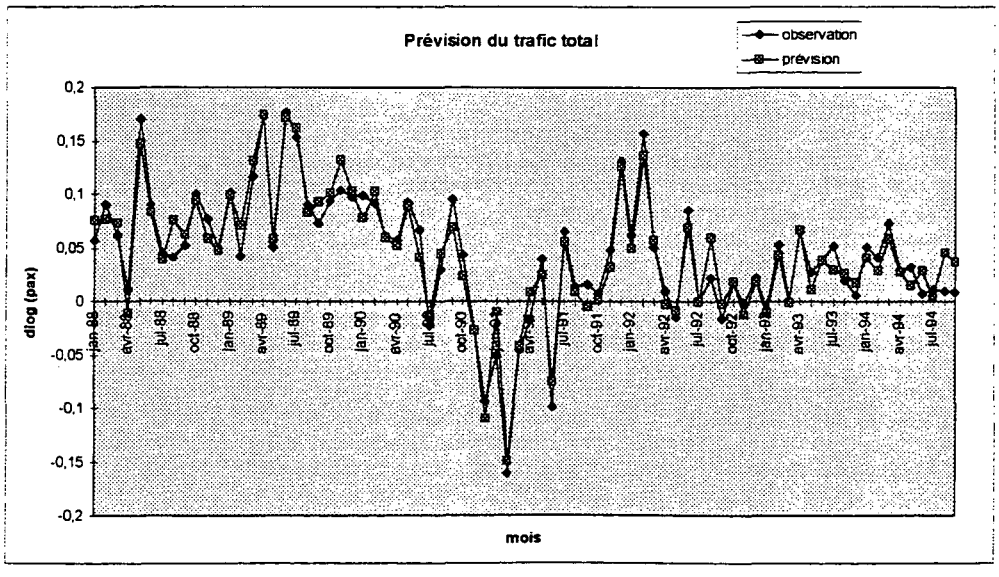
Model fitting results for: total

Independent variable	coefficient	std. error	t-value	sig.level
CONSTANT	0.020988	0.002631	7.9765	0.0000
BASTIA NICE	0.091681	0.022899	4.0036	0.0001
BIARRITZ ORLY	0.304536	0.022571	13.4926	0.0000
BORDEAUX CDG	0.061049	0.009592	6.3643	0.0000
LYON TOULOUSE	0.173732	0.025479	6.8187	0.0000
ORLY RENNES	0.135869	0.010007	13.5770	0.0000

R-SQ. (ADJ.) = 0.9196 SE= 0.016768 MAE= 0.012936 DurbWat= 1.841
78 observations fitted, forecast(s) computed for 0 missing val. of dep. var.

Les séries chronologiques du trafic observé et calculé sont visualisées dans la figure 10.

fig 10 : trafic observé et trafic modélisé



III-2 Validation du modèle

Voici les résultats de l'analyse de la variance :

Analysis of Variance for the Full Regression

Source	Sum of Squares	DF	Mean Square	F-Ratio	P-value
Model	0.248939	5	0.0497877	177.073	.0000
Error	0.0202442	72	0.000281170		
Total (Corr.)	0.269183	77			

R-squared = 0.924794 Stnd. error of est. = 0.0167681
R-squared (Adj. for d.f.) = 0.919571 Durbin-Watson statistic = 1.84108

Quelques rappels pour l'interprétation :

coefficient = $\hat{a_i}$ pour i=1,...,5

standard error = s_i estimation de l'écart type de $\hat{a_i}$

t-value = $t_i = \frac{\hat{a_i}}{s_i}$ suit une loi de Student à 72 degrés de liberté (n-k-1). Au seuil de 5%, la ligne est significative si $|t_i| \geq 1.994$

significance level = Prob$(|T(72)| \geq |t_i|)$

SE = Standard error of estimation = $\hat{\sigma}$ estimation de l'écart type de ε

DurbWat = Durbin-Watson statistic = DW = statistique d'un test permettant de détecter une autocorrélation des erreurs d'ordre 1. Au seuil de 5%, pour 78 observations et 5 variables + 1 terme constant, on accepte l'hypothèse que les erreurs ne sont pas corrélées pour $DW \in [1.77, 2.23]$.

F-ratio = statistique du test global de significativité

$$F = \frac{\sum_{t=1}^{n} \left(\hat{Y_t} - \bar{Y}\right)^2 / k}{\sum_{t=1}^{n} e_t^2 / (n-k-1)} = \frac{SumOfSquares(Model)/DF(Model)}{SumOfSquares(Error)/DF(Error)}$$

Le test est :

$H_0 : a_0 = a_1 = ... = a_k = 0$

contre H_1 : au moins un $a_i \neq 0$

Si H_0 est vraie, F suit une loi de Fisher-Snedecor à k et n-k-1 degrés de liberté. On rejette H_0 avec un risque d'erreur α si :

$F \geq F_{1-\alpha}(k, n-k-1)$

Avec 78 observations et 5 variables, le modèle est globalement significatif si $F \geq 2.75$.

Le modèle est donc globalement hautement significatif, chaque ligne l'est aussi, et le test de Durbin Watson est bon.

Le R^2 ajusté de 0.9196 indique que le pouvoir explicatif du modèle est satisfaisant.

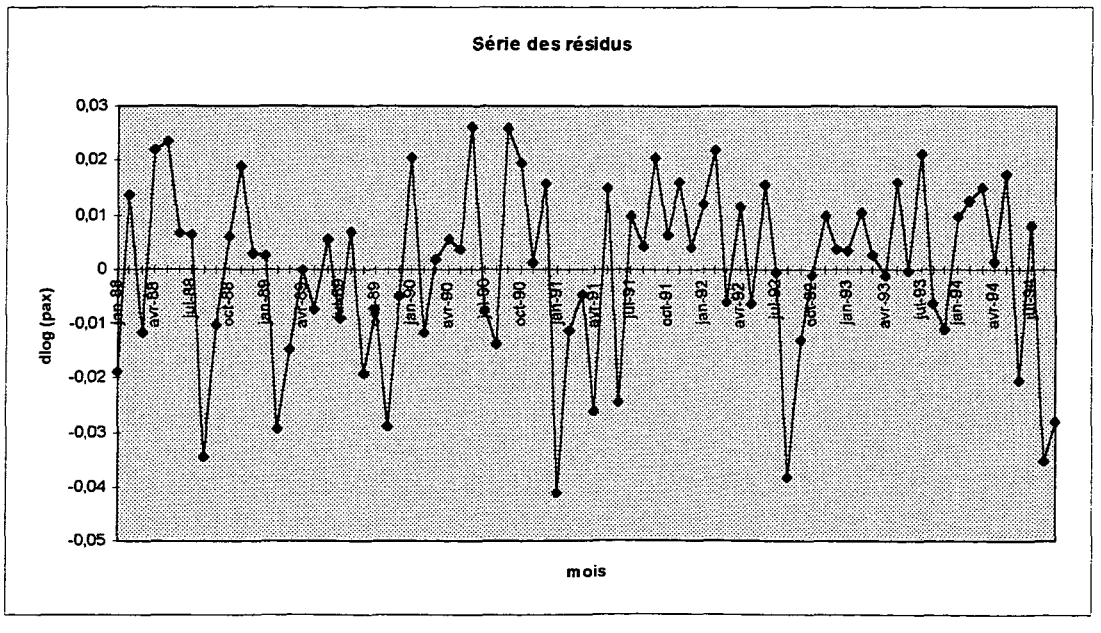
La figure 11 présente la matrice de corrélation des variables.

fig 11 : tableau des corrélations entre les variables

	BASTIA NICE	BIARRITZ ORLY	BORDEAUX CDG	LYON TOULOUSE	ORLY RENNES
TRAFIC TOTAL	0,27	0,79	0,57	0,37	0,53
BASTIA NICE	1	0,16	0,03	0,26	-0,03
BIARRITZ ORLY		1	0,52	0,19	0,18
BORDEAUX CDG			1	0,12	0,04
LYON TOULOUSE				1	-0,05
ORLY RENNES					1

Il ne semble pas y avoir de risques graves de multicolinéarité des variables. Néanmoins la corrélation de 0.52 interdit l'interprétation des coefficients en termes d'élasticité (à cause des effets de cache), ce qui n'est pas crucial dans notre cas.

fig 12 : Série des erreurs de prévision



La figure 12 permet de visualiser les résidus. On constate que les mois d'août 1988, janvier 1991, août 1992 et septembre 1994 ont été mal reconstitués par le modèle.

Les mois remarquables par leur résidu "studentisé", leur levier[6] ou leur DFITS[7]ont été regroupés dans le tableau des "flagged observations".

Flagged Observations for total

Mois	Stnd. Residual	Leverage	Mahalanobis Dist.	DFITS
janvier 1988	-1.25480	0.18309	16.0464	**-0.59404**
mai 1988	1.56201	0.16803	14.3627	**0.70198**
août 1988	**-2.16841**	0.05921	3.79573	**-0.54397**
mai 1989	-0.52254	**0.29105**	30.2143	-0.33481
novembre 1989	-1.84322	0.09876	7.34100	**-0.61016**
juin 1990	1.69361	0.12411	9.78184	**0.63752**
janvier 1991	**-2.69117**	0.09750	7.22314	**-0.88452**
août 1992	**-2.36284**	0.01637	0.27748	-0.30478
septembre 1994	**-2.18935**	0.02704	1.12490	-0.36496

Number of flagged observations (residual, leverage or DFITS) = 9

Résidu "studentisé"[8]:
On peut montrer que le résidu $e_t = Y_t - \hat{Y_t}$ est une réalisation d'une variable aléatoire suivant une loi normale $N\left(0, \sigma\sqrt{1-h_t}\right)$ où h_t est le levier du mois t. Ainsi, plus un mois est une observation atypique (fort levier), plus il attire le modèle estimé (faible erreur).
Pour mesurer l'importance du résidu e_t on utilise le résidu "studentisé" :

$$RS_t = \frac{e_t}{\hat{\sigma}\sqrt{1-h_t}}$$

Les valeurs considérées comme grandes pour chaque critère (résidu studentisé, levier et DFITS) sont en gras dans le tableau ci-dessus.

On retrouve que les mois d'août 1988, janvier 1991, août 1992 et septembre 1994 sont mal reconstitués par le modèle.

Les mois de janvier 1988, mai 1988, août 1988, novembre 1989, juin 1990 et janvier 1991 peuvent être considérés comme influents.

On constate que le mois atypique mai 1989 (fort levier) n'a pas une influence particulière sur l'estimation du modèle (DFITS normal).

Au vu de la figure 12, les résidus ont bien l'allure de réalisations d'un bruit blanc. Il convient néanmoins d'approfondir leur étude par les graphiques de la figure 13, où l'on ne détecte pas d'hétéroscédasticité. La figure 14, qui visualise les corrélogrammes simples et partiels des résidus, montre qu'il ne semble pas y avoir d'autocorrélation des erreurs (hormis une corrélation d'ordre 10 imputable vraisemblablement à une valeur aberrante).

Le modèle est donc satisfaisant sur le plan statistique.

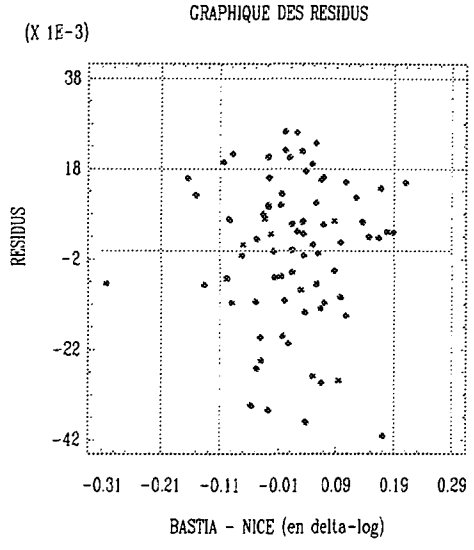
---

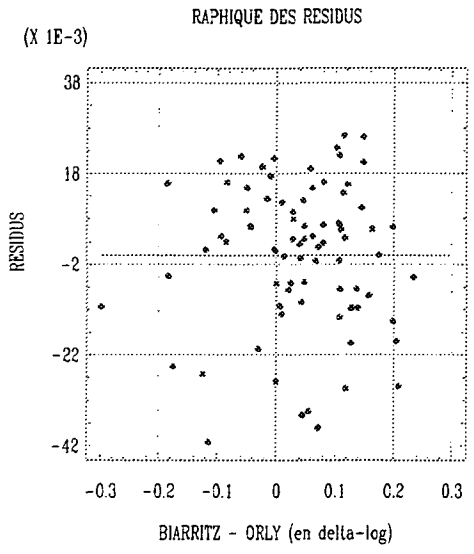
6 voir page 26

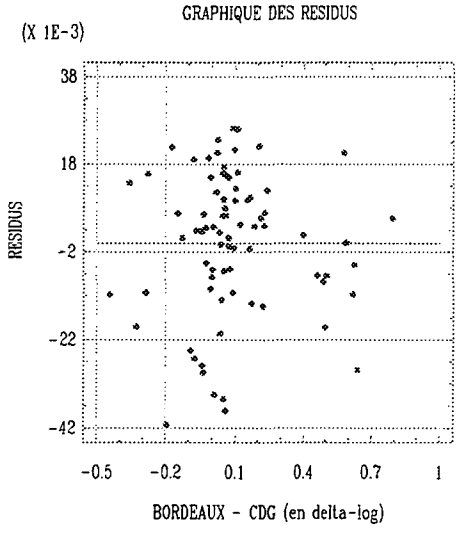
7 voir page 25

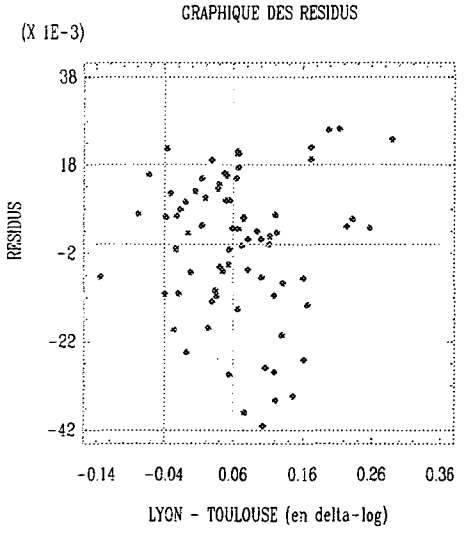
8 tiré de *Méthodes statistiques en Gestion* (p95) de Michel Tenenhaus

fig 13 : Les résidus fonctions des lignes









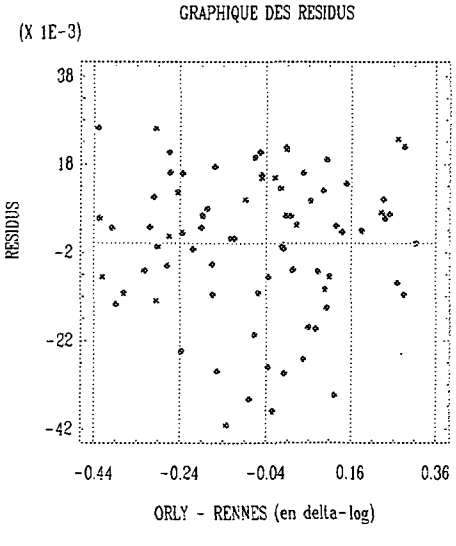
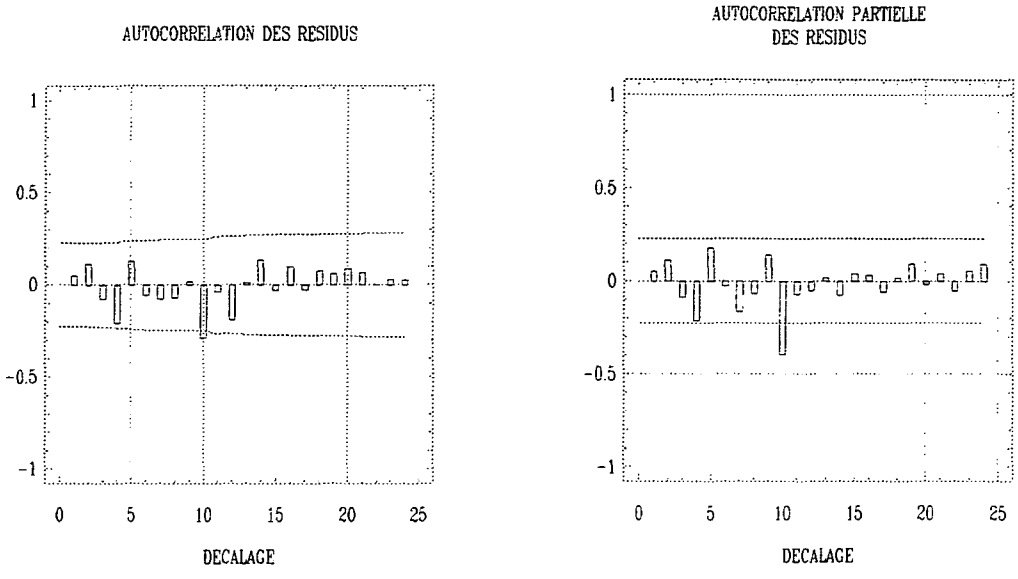


fig 14 : Corrélogrammes des résidus



III-3 Résumé du modèle

Le modèle estimé s'écrit :

DOMESTIQUE-8RADCONCU= 0.021 + 0.092×BASTIA NICE+0.305×BIARRITZ ORLY
+0.061×BORDEAUX CDG+0.174×LYON TOULOUSE
+0.136×ORLY RENNES

Pour passer de la valeur en delta-log (D_t) au trafic en nombre de passagers (P_t), il faut utiliser la formule :

$P_t = \exp(D_t)P_{t-12}$

PARTIE III :

CONSTRUCTION D'UN OUTIL DE PREVISION A COURT TERME

Maintenant que nous disposons d'un outil de suivi du trafic total hors concurrence, c'est-à-dire un outil de "prévision" dans le passé, à partir du trafic *connu* de cinq lignes aériennes, nous aimerions pouvoir prévoir les valeurs futures du trafic.

Nous pouvons faire de la prévision directement à partir de la série de trafic total, ou bien prévoir le trafic des cinq lignes explicatives, et utiliser la relation de régression déterminée dans la partie II. Il sera intéressant de comparer les deux modes de prévision.

Dans les deux cas nous devons faire de la prévision de séries temporelles. Cette prédiction se fera à partir de la connaissance des valeurs passées de la série elle-même, et non à partir de modèles explicatifs, c'est-à-dire de séries temporelles externes. Il ne pourra donc s'agir que de prévision à court terme (quelques mois).

Nous allons commencer par étudier, à tour de rôle, la série du trafic total, et les séries de trafic des cinq lignes précédemment retenues.

I - Prévisions par la méthode de Box et Jenkins

La méthodologie de Box et Jenkins consiste à choisir dans la vaste classe des modèles SARIMA le modèle représentant au mieux la série étudiée.

Les modèles ARMA[9]

Pour pouvoir tirer des enseignements de l'observation du passé d'un processus, notamment pour pouvoir faire de la prévision, il est pratique de supposer que la loi de probabilité est stable au cours du temps, c'est-à-dire que le modèle est *stationnaire*. Ainsi, pour un processus discret stationnaire (x_t), tous les x_t ont même moyenne μ, même variance σ^2, et les autocorrélations $\rho_k = corr(x_t, x_{t-k})$ sont indépendantes de l'instant t pour tout décalage k.
On supposera pour commencer que le processus aléatoire considéré est stationnaire et centré (μ=0).
On peut alors l'obtenir à partir d'une suite (ε_t) d'aléas indépendants et de même loi, qualifiés de chocs aléatoires, sous forme d'une combinaison linéaire infinie des chocs qui se sont produits dans le passé, d'où une expression de la forme :

$$x_t = \sum_{j=0}^{\infty} \psi_j \varepsilon_{t-j} \text{ avec } \psi_0 = 1 \text{ et } \sum_{j=0}^{\infty} |\psi_j| \leq \infty$$

Sous certaines conditions de régularité, un tel processus est inversible et peut adopter une écriture de la forme :

$$x_t = \varepsilon_t + \sum_{j=1}^{\infty} \phi_j x_{t-j} = \varepsilon_t + \hat{x}_t$$

ce qui indique que x_t est la somme du dernier choc ε_t et d'une fonction linéaire de son passé. L'aléa ε_t s'interprète donc comme *l'innovation* du processus par rapport à son passé, dont il est indépendant, et $\hat{x}_t$ qui est la régression linéaire de x_t sur son passé, comme la meilleure prévision de x_t connaissant son passé $x_{t-1}, x_{t-2}, \ldots$
Pour obtenir des modèles comportant un nombre fini de paramètres, on tronque les modèles précédents. On obtient ainsi :
le *processus moyenne mobile d'ordre q, MA(q)* :

$$x_t = \varepsilon_t - \sum_{j=1}^{q} \theta_j \varepsilon_{t-j} \text{ où } \theta_q \neq 0$$

et le *processus autorégresif d'ordre p, AR(p)* :

$$x_t = \varepsilon_t + \sum_{j=1}^{p} \phi_j x_{t-j} \text{ où } \phi_p \neq 0$$

Notons B l'opérateur de décalage arrière défini par $Bx_t = x_{t-1}$. On peut alors reformuler l'écriture des processus MA et AR :
pour un MA(q) : $x_t = \theta(B)\varepsilon_t$ où $\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$. Le processus est inversible si les racines du polynome θ satisfont certaines conditions.
pour un AR(p) : $\phi(B)x_t = \varepsilon_t$ où $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$. Le processus est stationnaire si les racines du polynome ϕ satisfont certaines conditions.
Sous ces conditions, les modèles AR(p) et MA(q) ont des propriétés duales.

---

9 inspiré de *Analyse et prévision des séries chronologiques*, de D. Bosq et J-P. Lecoutre

Ainsi, l'*autocorrélation* ρ_k d'un AR(p), c'est-à-dire le coefficient de corrélation entre x_t et x_{t-k}, qui ne dépend pas de t, tend vers 0 (quand $k \to +\infty$) à vitesse exponentielle, alors que, pour un MA(q), on a $\rho_k = 0$ dès que k>q.
En revanche, si on considère l'*autocorrélation partielle* Φ_{kk}, c'est-à-dire le coefficient de corrélation entre x_t et x_{t-k}, quand on a supprimé l'influence des variables $x_{t-1}, \ldots, x_{t-k-1}$ sur x_t et x_{t-k}, elle tend vers 0 à vitesse exponentielle pour un MA(q), alors que, pour un AR(p), on a $\Phi_{kk} = 0$ dès que k>p.
Une classe plus générale de processus est obtenue en considérant un modèle mixte, comportant simultanément une partie autorégressive et une partie moyenne mobile. Il s'agit du processus autorégressif - moyenne mobile d'ordre p,q, noté ARMA(p,q) et qui s'écrit :

$$\phi(B)x_t = \theta(B)\varepsilon_t$$

avec : $\phi(B) = 1 - \phi_1 B - \ldots - \phi_p B^p$

$\theta(B) = 1 - \theta_1 B - \ldots - \theta_q B^q$

Quand ce processus est inversible et stationnaire (sous certaines conditions portant sur les racines de θ et ϕ), ρ_k et ϕ_{kk} tendent vers 0 à vitesse exponentielle.

Extension aux processus ARIMA et SARIMA

Les modèles précédents ne comportant ni tendance ni saisonnalité, nous allons voir comment il est possible malgré tout de prendre en compte ces différents éléments.

Processus non saisonnier :
Si le processus (x_t) n'est pas centré, il suffit de remplacer x_t par $x_t - \mu$.
Si la moyenne μ_t de la variable x_t (la tendance) est un polynome en t de degré d, le processus $((1-B)^d x_t)$ des différences d'ordre d du processus (x_t) a une moyenne constante. Un processus non stationnaire qui peut être transformé en un processus ARMA(p,q) par différentiation d'ordre d est appelé processus ARIMA(p,d,q).

Processus saisonnier :
Supposons des données mensuelles, et s=12 l'ordre de la saisonnalité. Pour chaque mois de l'année, on possède alors une série de données annuelles, non saisonnières, et on peut construire un modèle ARIMA(P,D,Q). Supposons que ce soit le modèle pour les différents mois :

$$\Phi(B^s)(1-B^s)^D x_t = \Theta(B^s)u_t$$

avec $\Phi(B^s) = 1 - \Phi_1 B^s - \ldots - \Phi_P B^{sP}$

$\Theta(B^s) = 1 - \Theta_1 B^s - \ldots - \Theta_Q B^{sQ}$

Les processus $(u_t), (u_{2t}), \ldots, (u_{12t})$ sont chacun des bruits blancs. Mais on peut penser que les u_t sont corrélés entre eux. On suppose donc que les u_t suivent un modèle ARIMA(p,d,q) :

$$\phi(B)(1-B)^d u_t = \theta(B)\varepsilon_t$$

où cette fois les ε_t forment un bruit blanc.
En combinant les deux modèles ARIMA, on obtient le modèle SARIMA(p,d,q)(P,D,Q)$_s$:

$$\phi(B)\Phi(B^s)(1-B)^d(1-B^s)^D x_t = \theta(B)\Theta(B^s)\varepsilon_t$$

Identification, estimation, validation et prévision

Etant donnée une série temporelle observée $x_1, \ldots, x_n$,on cherche à *identifier* le modèle ARIMA(p,d,q) qui s'adapte le mieux à ces observations.
Des transformations diverses (centrage, différentiation...) permettent d'abord de se ramener à un ajustement par un modèle stationnaire ARMA(p,q).
Ensuite, on construit des estimateurs empiriques r_k et $\hat{\phi}_{kk}$ de ρ_k et ϕ_{kk}.
On utilise alors les propriétés théoriques indiquées au paragraphe précédent :

si $\hat{\phi}_{kk}$ s'annule au delà d'une certaine valeur p, on choisit un AR(p);
si r_k s'annule au delà d'une certaine valeur q, on choisit un MA(q);
sinon il faut essyer un modèle ARMA(p,q) plus difficile à identifier.

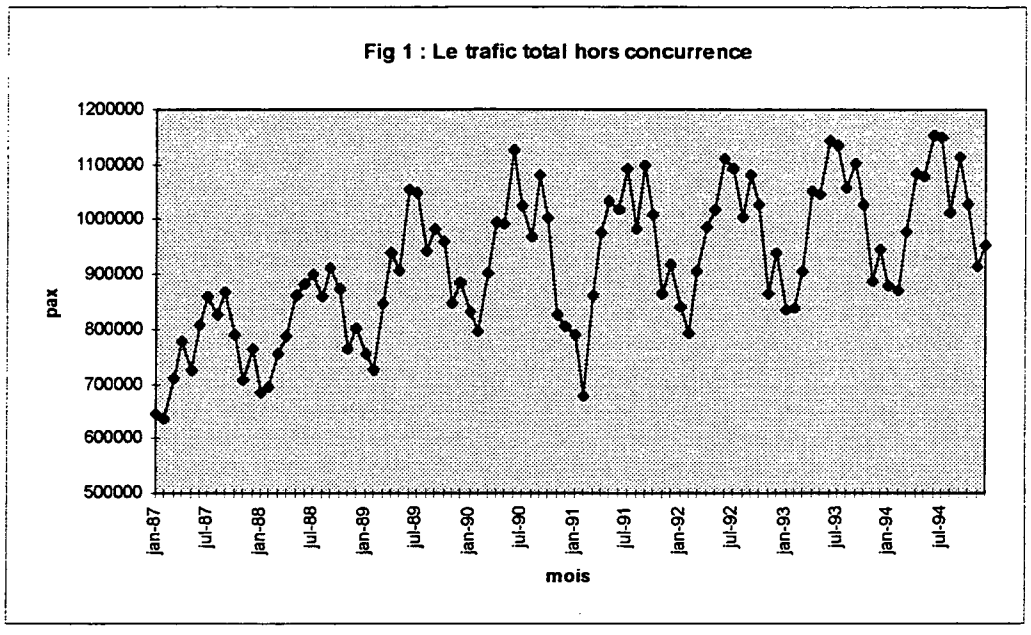
On passe alors aux phases d'*estimation* et de *validation* du modèle (l'hypothèse fondamentale à tester est l'indépendance des innovations ε_t, ce qui se fait par l'examen des corrélogrammes des résidus et par le test que les anglo-saxons appellent "portmanteau test".)

Si le modèle est accepté, on peut passer à la *prévision*.

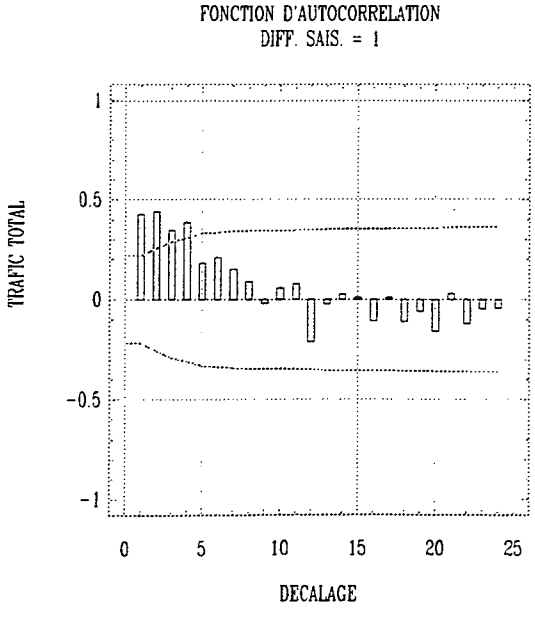
Pour éviter de biaiser la modélisation, nous avons décidé de remplacer le trafic des mois d'octobre et novembre 1993, anormalement faible à cause de grèves chez Air France, par la demi-somme des trafics des mêmes mois des années 1992 et 1994. Rappelons qu'une correction du même type a déjà été faite pour les mois de mars et avril 1989 sur les lignes vers la Corse, ce qui s'est répercuté sur le trafic total.

I-1- Le trafic total hors concurrence

La série du trafic total hors concurrence est représentée sur la figure 1.



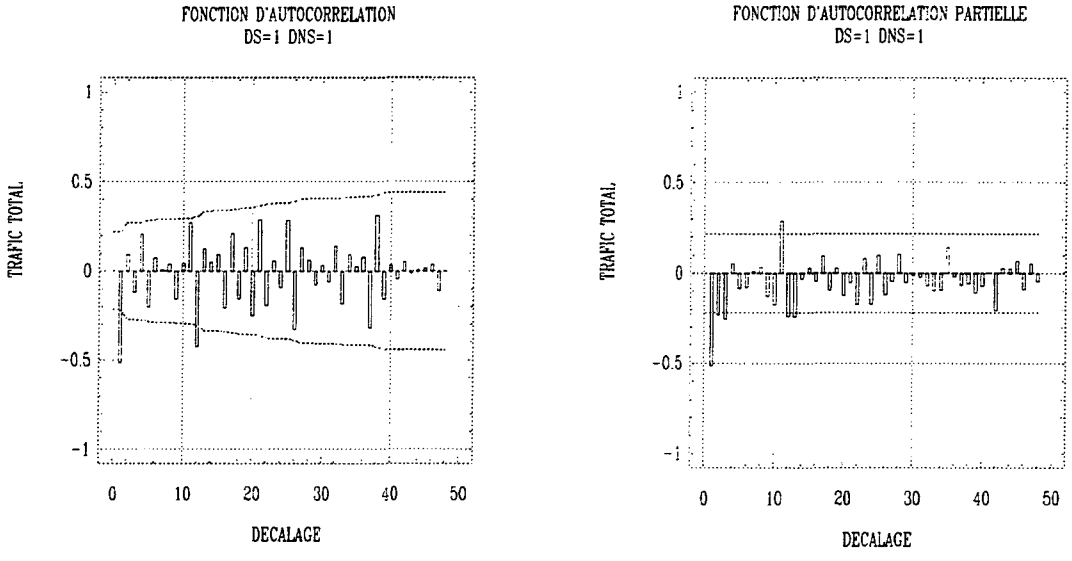
Elle possède bien sûr une composante saisonnière que nous stabilisons par une différentiation saisonnière d'ordre 1.



Le corrélogramme montre que la série n'est pas stationnaire en tendance. Nous la stabilisons alors par une différentiation régulière d'ordre 1.

La modélisation de la série résultante commence par l'étude des corrélogramme et corrélogramme partiel estimés. Ils sont donnés dans la figure 2. Les limites de signification y sont données en pointillé.

Fig 2 : Corrélogramme et corrélogramme partiel de la série du trafic total (différences saisonnière et régulière d'ordre 1)



En non saisonnier la décroissance des autocorrélations après r_1 est plus rapide que celle des autocorrélations partielles.

C'est la même chose en saisonnier après r_{12}.

Tout ceci nous conduit à essayer le modèle SARIMA$(0,1,1)(0,1,1)_{12}$.

Voilà les résultats des estimations :

Summary of Fitted Model for: total

Parameter	Estimate	Stnd.error	T-value	P-value
MA (1)	.59038	.08540	6.91275	.00000
SMA(12)	.86893	.04131	21.03291	.00000
MEAN	-600.24673	680.95823	-.88147	.38070
CONSTANT	-600.24673			

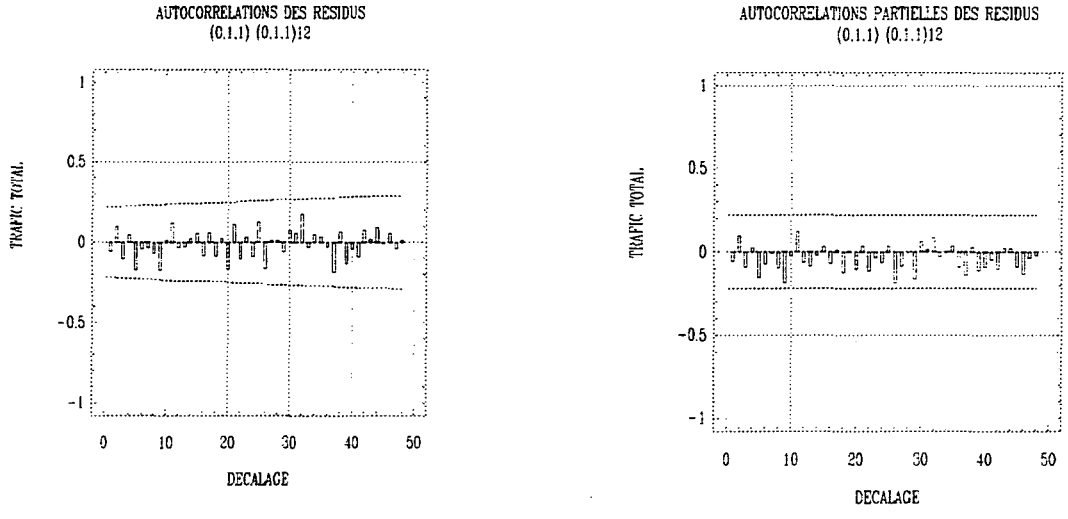
Model fitted to differences of order 1
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 1.34208E9 with 80 degrees of freedom.
Estimated white noise standard deviation (std err) = 36634.4
Chi-square test statistic on first 20 residual autocorrelations = 12.9527
with probability of a larger value given white noise = 0.739347
Backforecasting: yes Number of iterations performed: 8

Le modèle s'écrit :

$$(1-B)(1-B^{12})x_t = -600.25 + (1-0.59B)(1-0.87B^{12})\varepsilon_t$$

Le portmanteau test montre que l'hypothèse d'un bruit blanc pour les résidus est globalement acceptable (puisque $14.95 < \chi^2_{0.95}(18) = 28.9$). La figure 3 montre qu'il n'y a pas d'autocorrélation significative des résidus.

Fig 3 : Corrélogramme et corrélogramme partiel des résidus

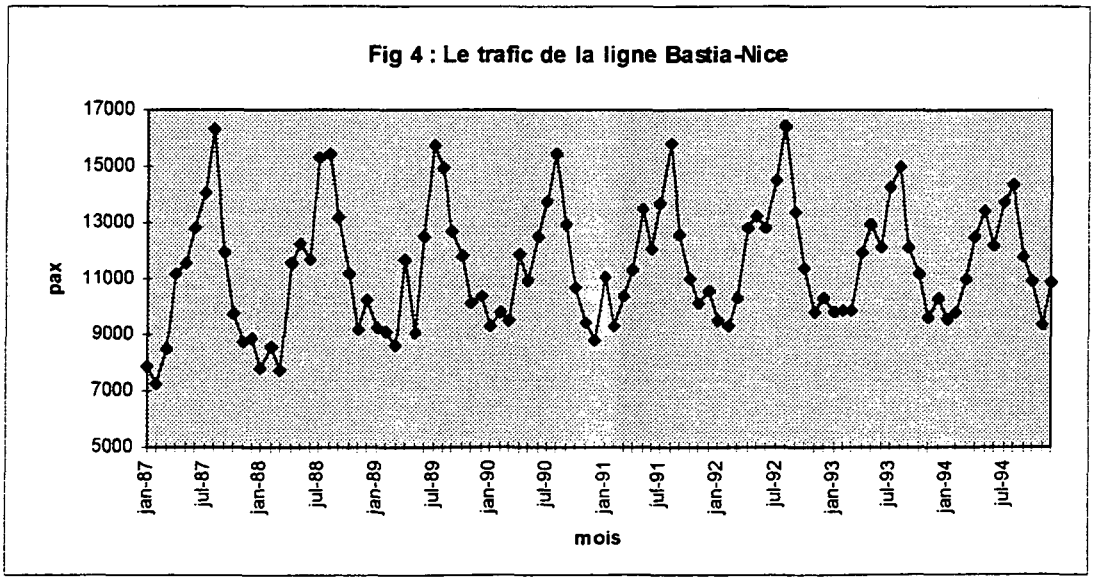


Le modèle est donc acceptable.

I-2- Bastia - Nice

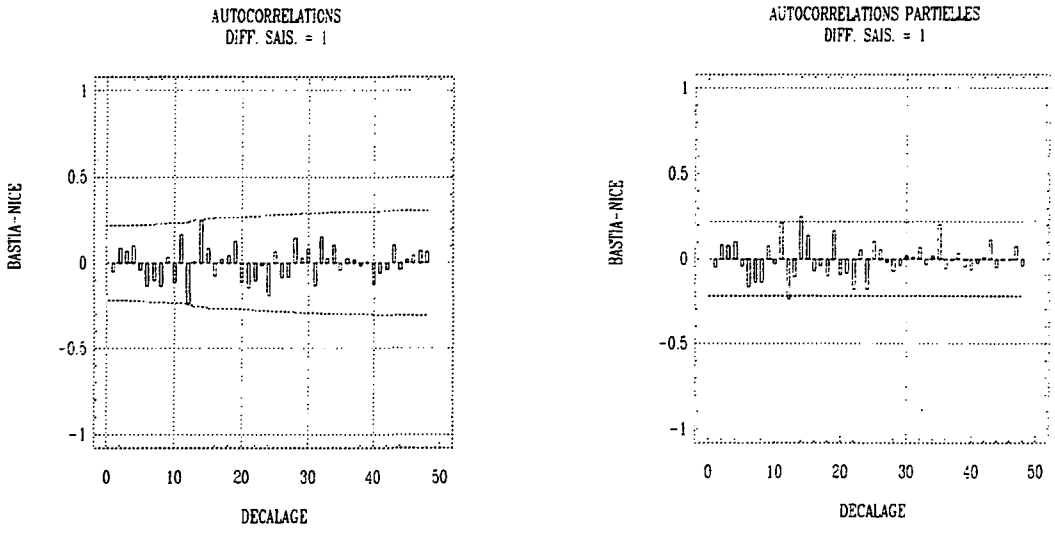
La série de trafic est représentée sur la figure 4.

Fig 4 : Le trafic de la ligne Bastia-Nice



Nous stabilisons la saisonnalité par une différentiation saisonnière d'ordre 1.

Fig 5 : Corrélogramme et corrélogramme partiel de la série de trafic de la ligne Bastia-Nice (différence saisonnière d'ordre 1)



En non saisonnier, aucune corrélation ni corrélation partielle n'est significative.

En saisonnier la décroissance brutale après r_{12} suggère un $MA(1)_{12}$.

Ceci nous conduit à essayer le modèle $SARIMA(0,0,0)(0,1,1)_{12}$.

Voici les résultats des estimations:

Summary of Fitted Model for: BASTIA-NICE

Parameter	Estimate	Stnd.error	T-value	P-value
SMA(12)	.46776	.10817	4.32418	.00004
MEAN	135.30011	59.08928	2.28976	.02461
CONSTANT	135.30011			

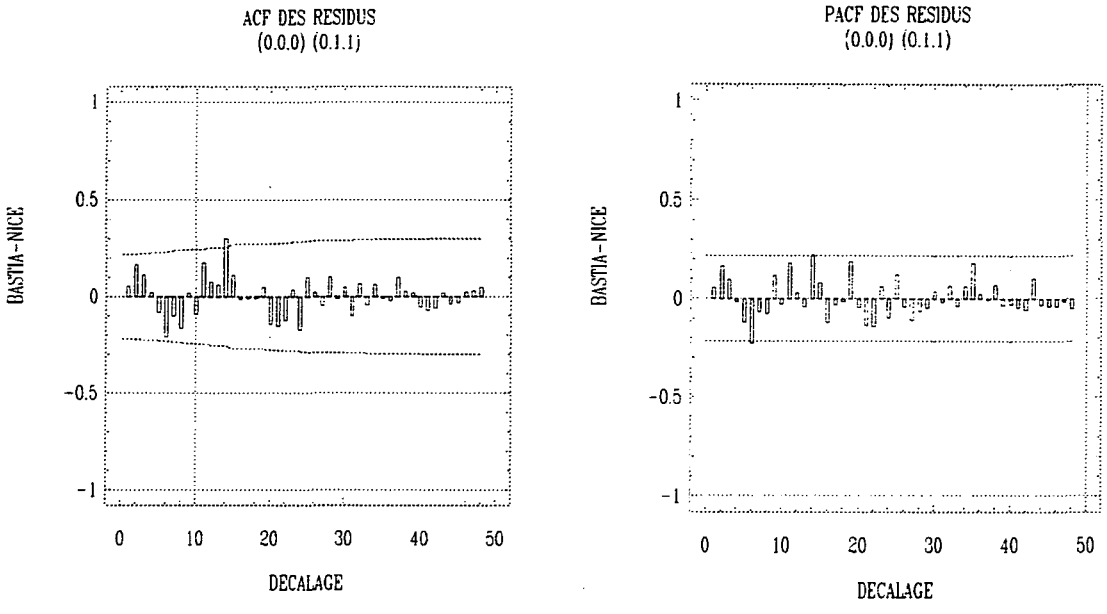
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 771695 with 82 degrees of freedom.
Estimated white noise standard deviation (std err) = 878.462
Chi-square test statistic on first 20 residual autocorrelations = 25.7595
with probability of a larger value given white noise = 0.105385
Backforecasting: no Number of iterations performed: 5

Le modèle s'écrit :

$$(1-B^{12})x_t = 135.30 + (1-0.47B^{12})\varepsilon_t .$$

Le portmanteau test est bon $\left(25.78 < \chi^2_{0.95}(19) = 30.1\right)$. La figure 6 montre qu'il n'y a pas d'autocorrélation significative des résidus (hormis une corrélation d'ordre 14 qui n'a pas de réalité économique).

Fig 6 : Corrélogramme et corrélogramme partiel des résidus

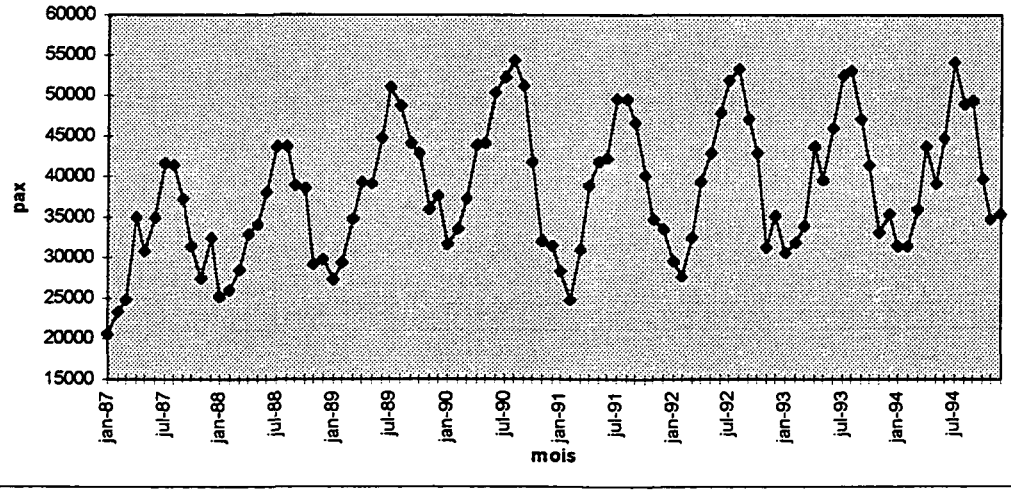


Le modèle est donc acceptable.

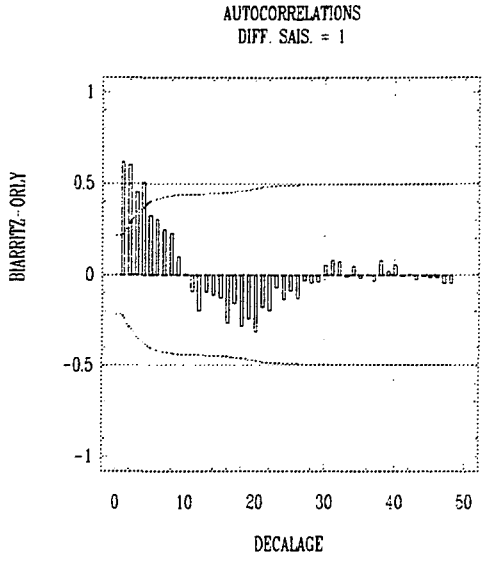
I-3- Biarritz - Orly

La série de trafic est représentée sur la figure 7.

Fig 7 : Le trafic de la ligne Biarritz-Orly

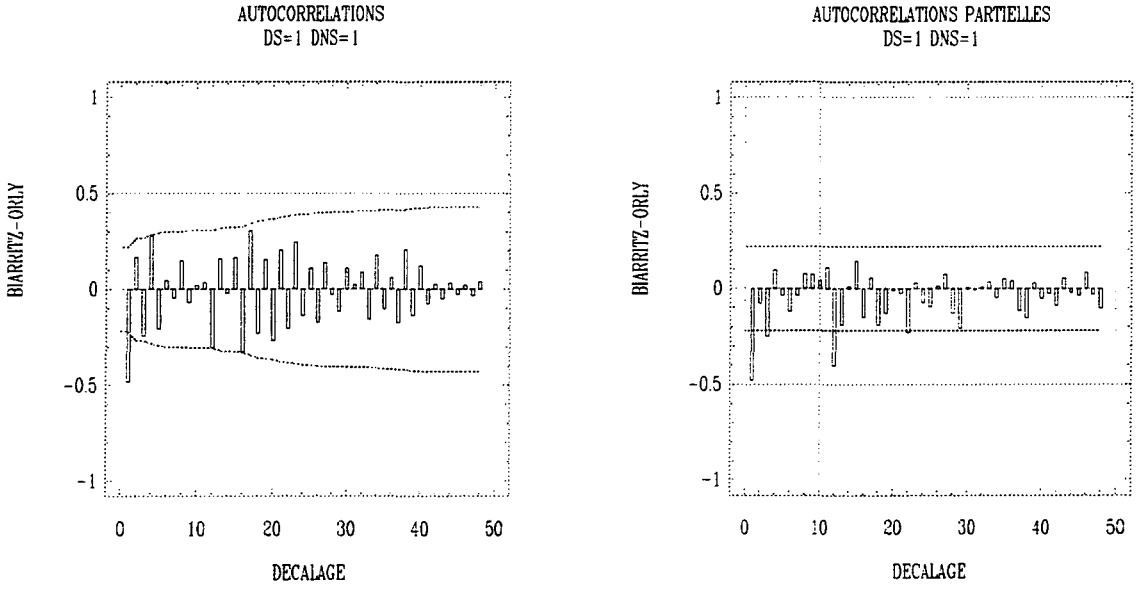


Nous stabilisons la saisonnalité par une différentiation saisonnière d'ordre 1.



Le corrélogramme montre que la série n'est pas stationnaire en tendance. Nous la stabilisons alors par une différentiation régulière d'ordre 1.

Fig 8 : Corrélogramme et corrélogramme partiel de la série de trafic de la ligne Biarritz-Orly (différences saisonnière et régulière d'ordre 1)



En saisonnier, la décroisance des autocorrélations partielles aprés $\hat{\phi}_{12.12}$ est plus rapide que celle des autocorrélations.

En non saisonnier la décroissance des autocorrélations partielles est plus rapide aprés $\hat{\phi}_{3.3}$ que celle des autocorrélations.

Ceci nous conduit à essayer le modèle SARIMA$(3,1,0)(1,1,0)_{12}$.

Voilà les résultats des estimations :

Summary of Fitted Model for: BIARRITZ-ORLY

Parameter	Estimate	Stnd.error	T-value	P-value
AR (1)	-.56840	.11055	-5.14159	.00000
AR (2)	-.20980	.12646	-1.65896	.10114
AR (3)	-.25609	.11355	-2.25542	.02691
SAR(12)	-.38882	.11119	-3.49684	.00078
MEAN	-41.83496	105.21725	-.39761	.69201
CONSTANT	-118.19520			

Model fitted to differences of order 1
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 6.75209E6 with 78 degrees of freedom.
Estimated white noise standard deviation (std err) = 2598.48
Chi-square test statistic on first 20 residual autocorrelations = 13.5053
with probability of a larger value given white noise = 0.56333
Backforecasting: no Number of iterations performed: 4

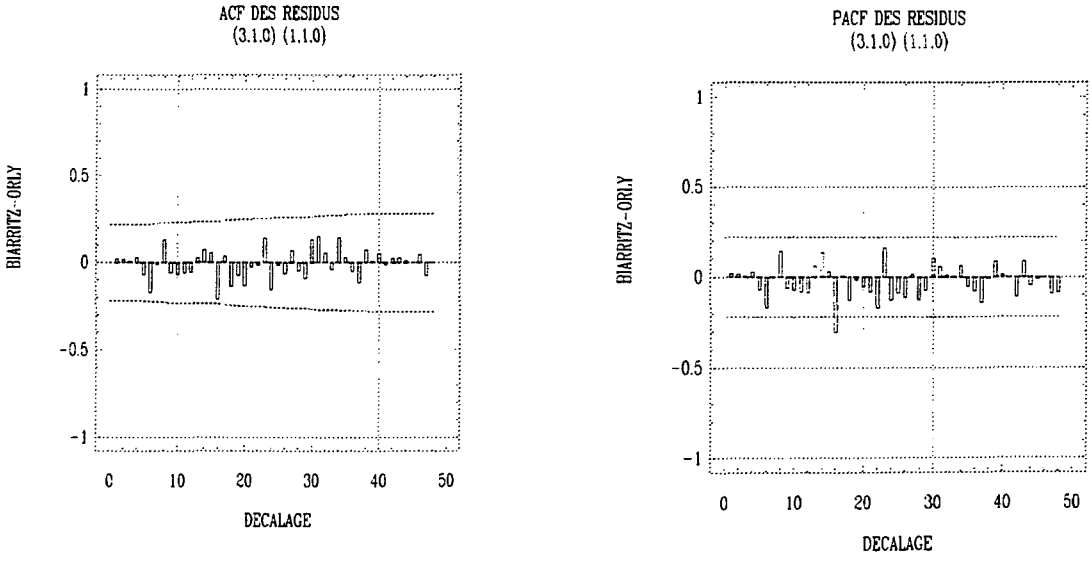
Le coefficient en AR(2) est non significativement différent de 0, comme on pouvait s'y attendre au vu du corrélogramme partiel.

Le modèle s'écrit :

$$(1+0.39B^{12})(1+0.57B)(1+0.21B^{2})(1+0.26B^{3})(1-B^{12})(1-B)x_t = -118.20+\varepsilon_t$$

Le portmanteau test est bon $\left(13.50 < \chi^2_{0.95}(16) = 26.3\right)$. La figure 9 montre qu'il n'y a pas d'autocorrélation significative des résidus (hormis une corrélation d'ordre 16 qui n'a pas de réalité économique).

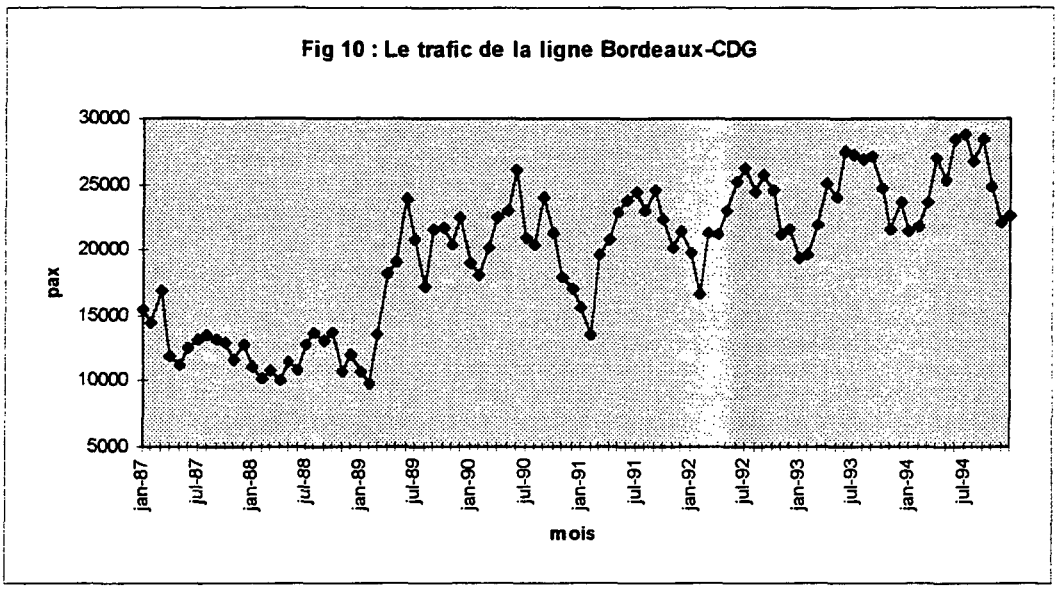
Fig 9 : Corrélogramme et corrélogramme partiel des résidus



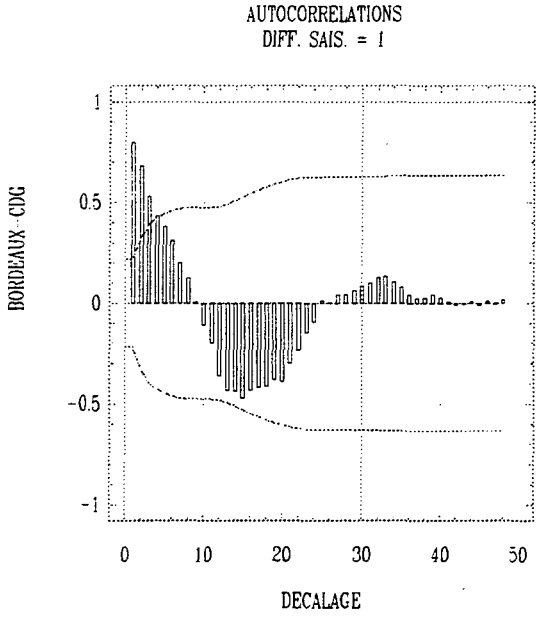
Le modèle est donc acceptable.

I-4- Bordeaux-Roissy Charles de Gaulle

La série de trafic est représentée sur la figure 10.



Nous stabilisons la saisonnalité par une différentiation saisonnière d'ordre 1.

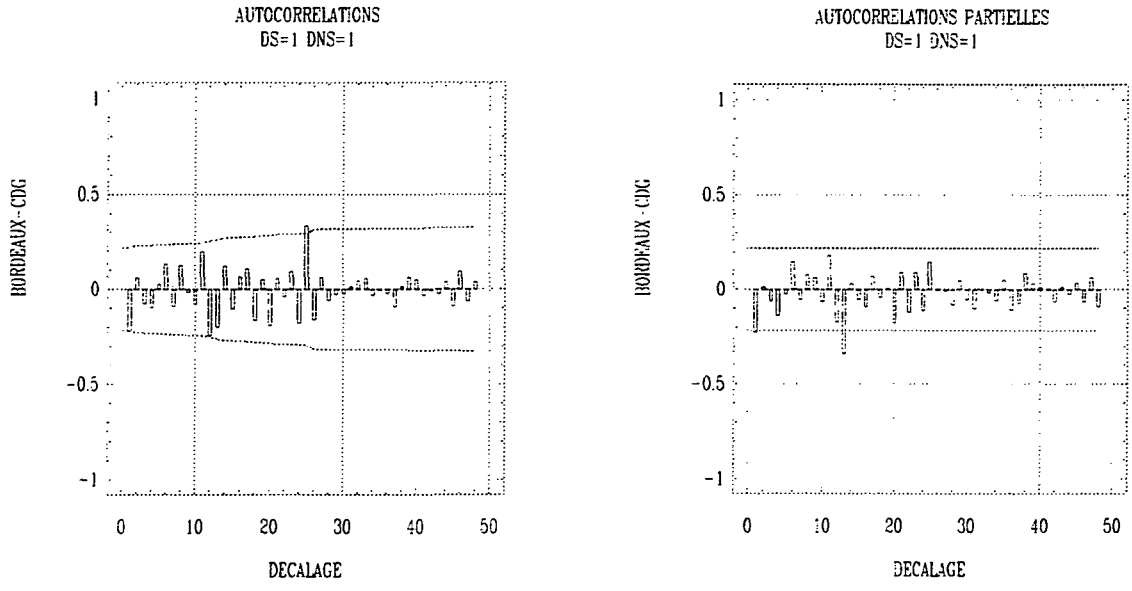


Le corrélogramme montre que la série n'est pas stationnaire en tendance. Nous la stabilisons alors par une différentiation régulière d'ordre 1.

En saisonnier, la décroissance des corrélations est plus rapide après r_{12} que celle des corrélations partielles. (voir fig 11)

En non saisonnier, l'interprétation est plus délicate. Plusieurs tentatives montrent que le modèle SARIMA$(0,1,1)(0,1,1)_{12}$ est le meilleur (plus petite variance résiduelle et meilleur portmanteau test).

Fig 11 : Corrélogramme et corrélogramme partiel de la série de trafic de la ligne Bordeaux-CDG (différences saisonnière et régulière d'ordre 1)



Voilà les résultats des estimations :

Summary of Fitted Model for: CPAX5.lig11

Parameter	Estimate	Stnd.error	T-value	P-value
MA (1)	.20869	.11178	1.86702	.06556
SMA(12)	.43356	.10920	3.97049	.00016
MEAN	42.40641	111.69053	.37968	.70519
CONSTANT	42.40641			

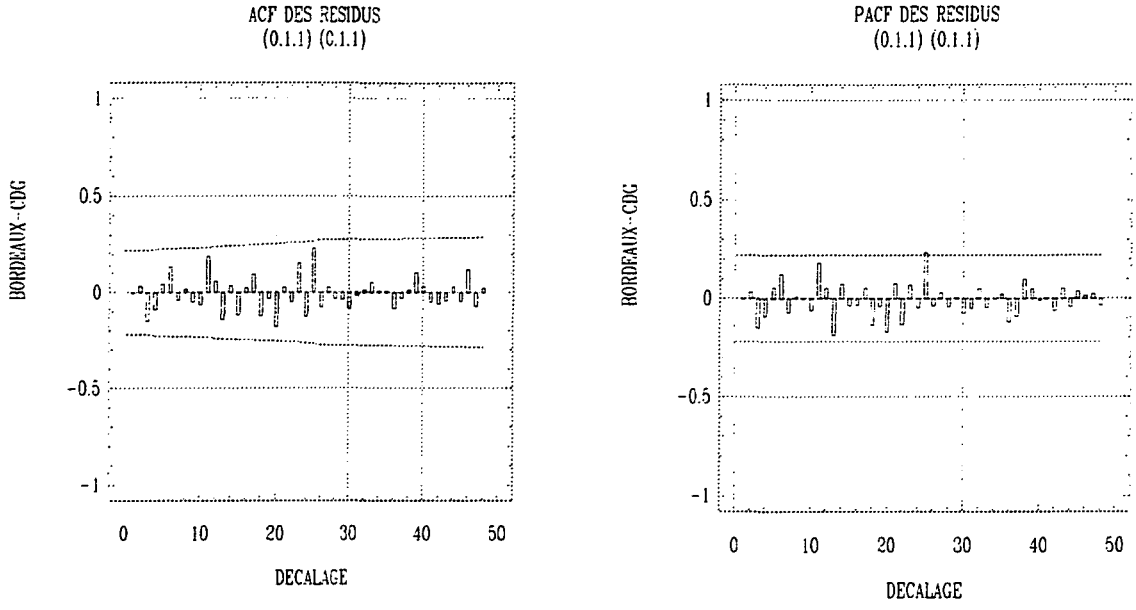
Model fitted to differences of order 1
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 3.92057E6 with 80 degrees of freedom.
Estimated white noise standard deviation (std err) = 1980.04
Chi-square test statistic on first 20 residual autocorrelations = 15.6081
with probability of a larger value given white noise = 0.551765
Backforecasting: no Number of iterations performed: 4

Le modèle s'écrit :

$$(1-B^{12})(1-B)x_t = 42.41 + (1-0.43B^{12})(1-0.21B)\varepsilon_t$$

Le portmanteau test est bon $\left(15.61 < \chi^2_{0,95}(18) = 28.9\right)$. La figure 12 montre qu'il n'y a pas d'autocorrélation significative des résidus.

Fig 12 : Corrélogramme et corrélogramme partiel des résidus

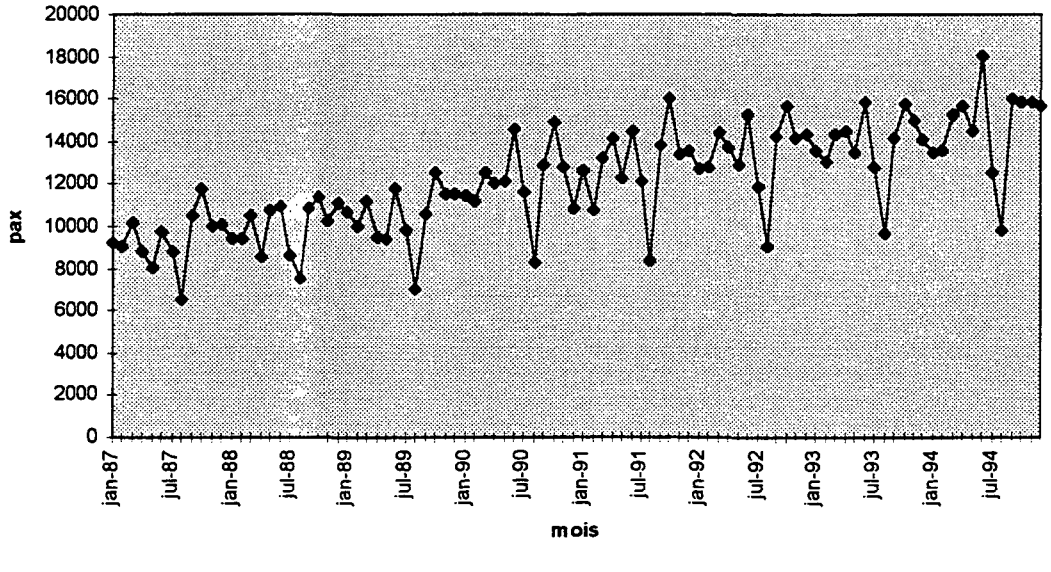


Le modèle est donc acceptable.

I-5- Lyon - Toulouse

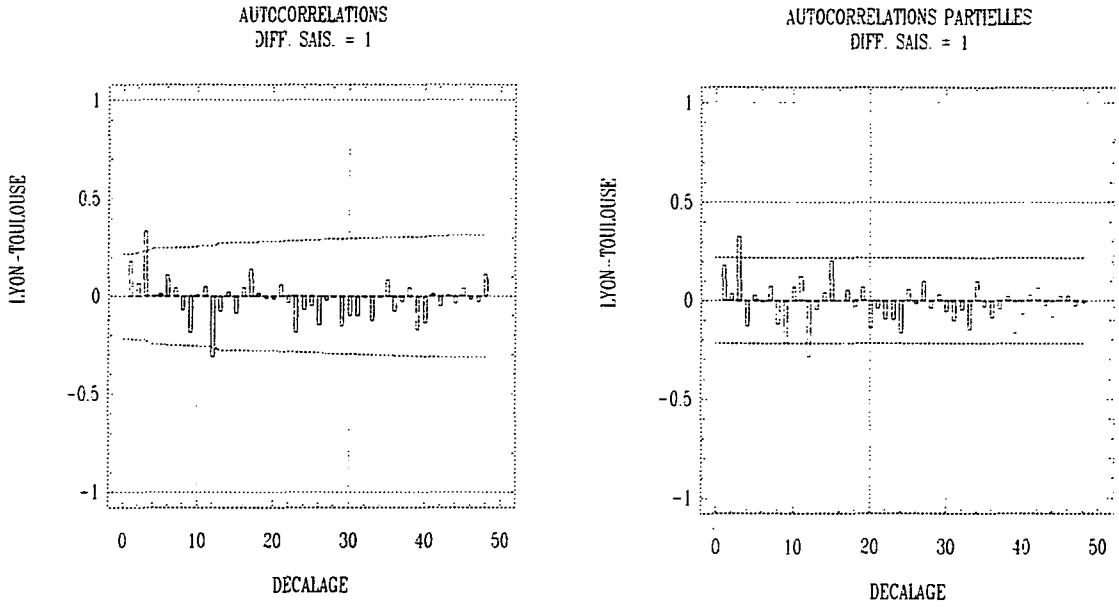
La série de trafic est représentée sur la figure 13.

Fig 13 : Le trafic de la ligne Lyon-Toulouse



Nous stabilisons la saisonnalité par une différentiation saisonnière d'ordre 1.

Fig 14 : Corrélogramme et corrélogramme partiel de la série de trafic de la ligne Lyon-Toulouse (différence saisonnière d'ordre 1)



En saisonnier, la décroissance des autocorrélations après r_{12} semble plus rapide que celle des autocorrélations partielles.

En non saisonnier, l'interprétation est plus délicate. Plusieurs tentatives montrent que le modèle SARIMA$(0,1,3)(0,1,1)_{12}$ est le meilleur (plus petite variance résiduelle et meilleur portmanteau test).

Voilà les résultats des estimations :

Summary of Fitted Model for: CPAX5.lig35

Parameter	Estimate	Stnd.error	T-value	P-value
MA (1)	-.28922	.11014	-2.62580	.01038
MA (2)	.01095	.11368	.09628	.92354
MA (3)	-.26384	.11054	-2.38683	.01939
SMA(12)	.47464	.11600	4.09172	.00010
MEAN	770.52805	76.88672	10.02160	.00000
CONSTANT	770.52805			

Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 577616 with 79 degrees of freedom.
Estimated white noise standard deviation (std err) = 760.011
Chi-square test statistic on first 20 residual autocorrelations = 11.8639
with probability of a larger value given white noise = 0.689304
Backforecasting: no Number of iterations performed: 4

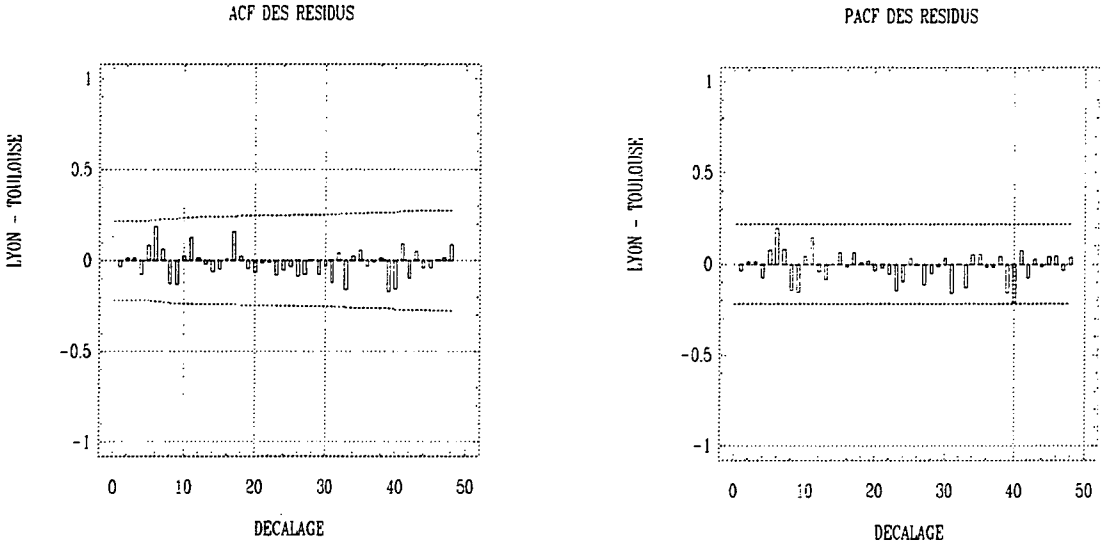
Le coefficient en MA(2) est non significativement différent de 0, comme on pouvait s'y attendre au vu du corrélogramme.

Le modèle s'écrit :

$$(1-B^{12})x_t = 770.53 + (1-0.47B^{12})(1+0.29B)(1-0.01B^2)(1+0.26B^3)\varepsilon_t$$

Le portmanteau test est bon $\left(11.86 < \chi^2_{0.95}(16) = 26.3\right)$. La figure 15 montre qu'il n'y a pas d'autocorrélation significative des résidus.

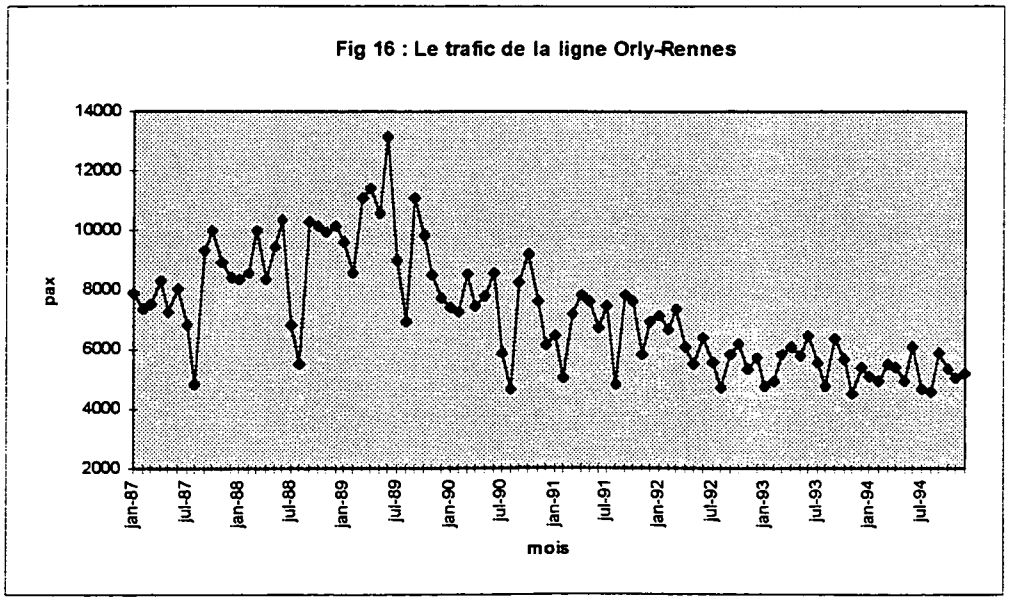
Fig 15 : Corrélogramme et corrélogramme partiel des résidus



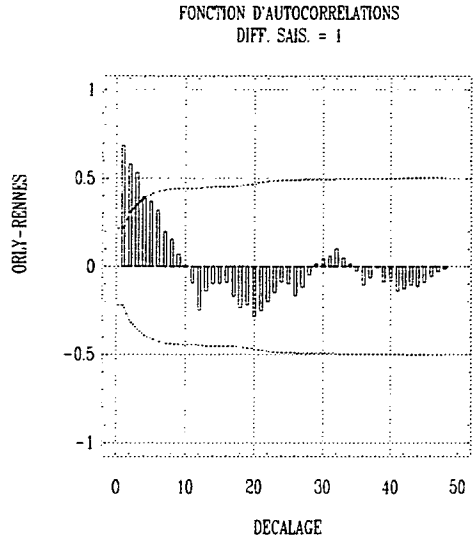
Le modèle est donc acceptable.

I-6- Orly - Rennes

La série de trafic est représentée sur la figure 16 :

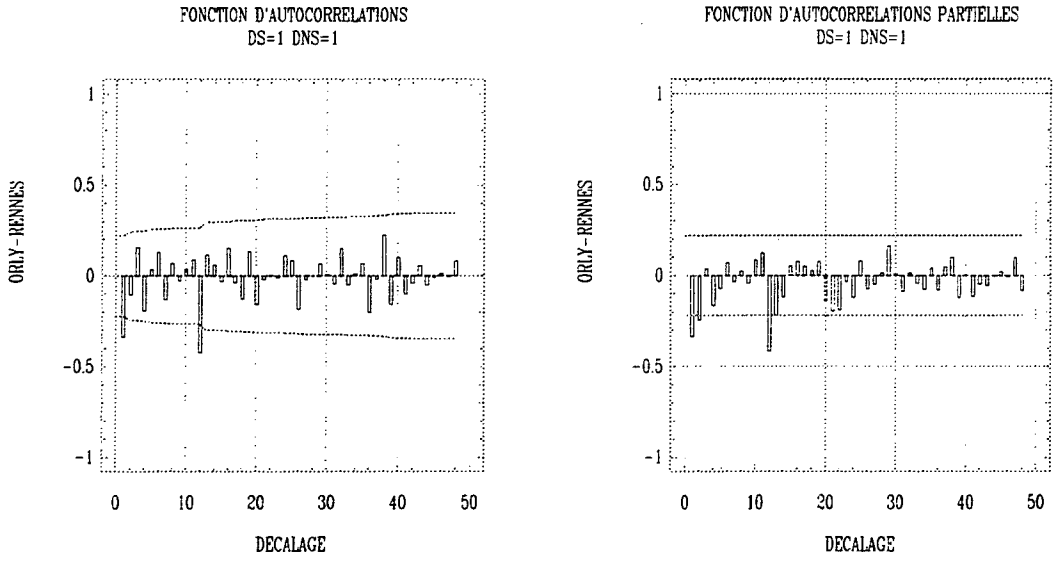


Nous stabilisons la saisonnalité par une différentiation saisonnière d'ordre 1.



Le corrélogramme montre que la série n'est pas stationnaire en tendance. Nous la stabilisons alors par une différentiation régulière d'ordre 1.

Fig 17 : Corrélogramme et corrélogramme partiel de la série de trafic de la ligne Orly-Rennes (différences saisonnière et régulière d'ordre 1)



En non saisonnier la décroissance des autocorrélations après r_1 est plus rapide que celle des autocorrélations partielles.

C'est la même chose en saisonnier après r_{12}.

Tout ceci nous conduit à essayer le modèle SARIMA$(0,1,1)(0,1,1)_{12}$.

Voilà les résultats des estimations :

Summary of Fitted Model for: ORLY-RENNES

Parameter	Estimate	Stnd.error	T-value	P-value
MA (1)	.49558	.09728	5.09409	.00000
SMA(12)	.57333	.10002	5.73218	.00000
MEAN	-12.06288	28.55482	-.42245	.67383
CONSTANT	-12.06288			

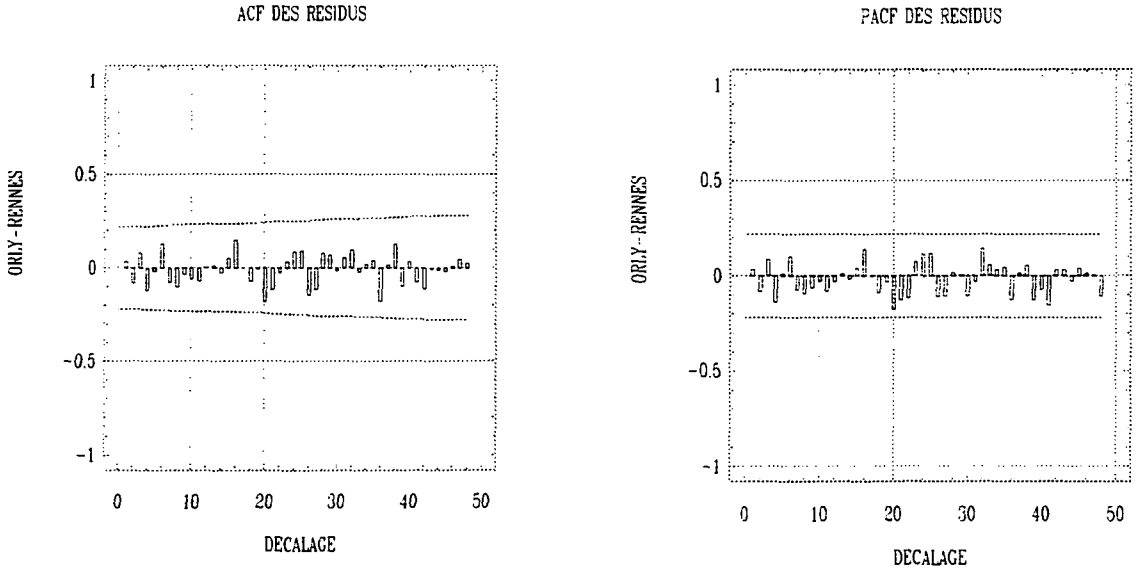
Model fitted to differences of order 1
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 941607 with 80 degrees of freedom.
Estimated white noise standard deviation (std err) = 970.364
Chi-square test statistic on first 20 residual autocorrelations = 11.1789
with probability of a larger value given white noise = 0.847112
Backforecasting: no Number of iterations performed: 4

Le modèle s'écrit :

$$(1-B^{12})(1-B)x_t = -12.06 + (1-0.57B^{12})(1-0.50B)\varepsilon_t$$

Le portmanteau test est bon $\left(11.18 < \chi^2_{0.95}(18) = 28.9\right)$. La figure 18 montre qu'il n'y a pas d'autocorrélation significative des résidus.

Fig 18 : Corrélogramme et corrélogramme partiel des résidus



Le modèle est donc acceptable.

II - Comparaison des deux méthodes de prévision

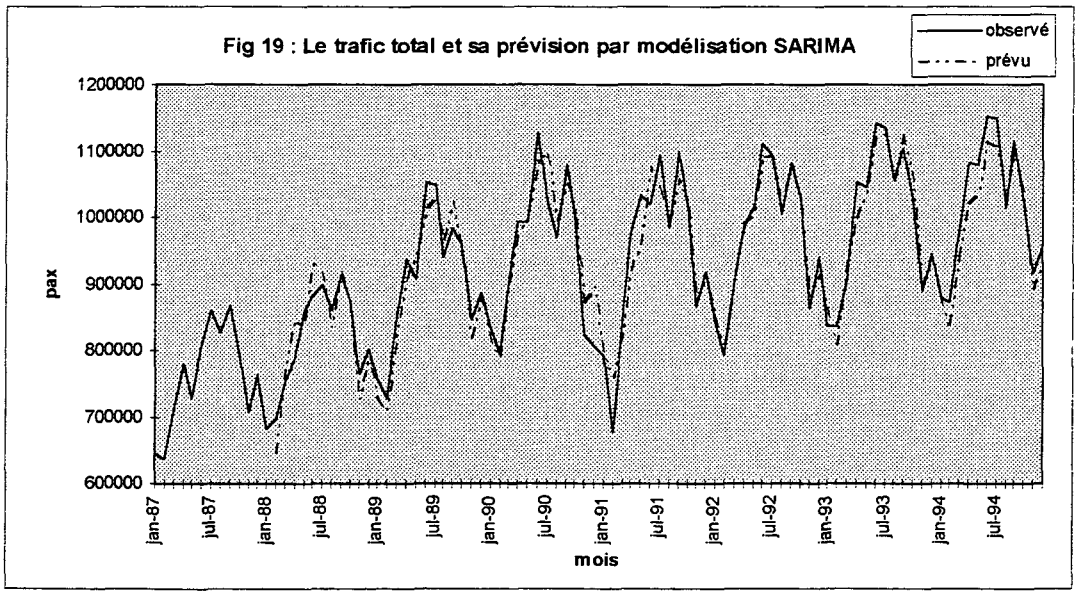
Rappelons les deux méthodes de prévision du trafic total hors concurrence :

- ✓ Utiliser la modélisation SARIMA du trafic total hors concurrence
- ✓ Combiner la relation de régression obtenue dans la partie II et les modélisations SARIMA des cinq lignes explicatives.

Pour pouvoir établir une comparaison entre ces deux méthodes, nous avons comparé les prévisions du trafic total hors concurrence pour l'année 1994, à partir des modèles calibrés sur la période janvier 1987 - décembre 1993.

II-1- Prévision directe

En conservant le modèle SARIMA $(0,1,1)(0,1,1)_{12}$ déterminé au I, nous avons recalculé les coefficients du modèle pour la période janvier 1987 - décembre 1993 (la structure même du modèle a peu de chances d'avoir changé). Nous avons alors établi des prévisions pour l'année 1994. Elles sont visualisées sur la figure 19.



Jusqu'à décembre 1993, il s'agit de prévision à un mois, faite le mois précédent. En revanche, pour l'année 1994, il s'agit de prévisions de 1 à 12 mois, effectuées en décembre 1993.

II-2- Prévision à l'aide de la régression

Il faut vérifier, dans un premier temps, que sur la période réduite 1987-1993, les cinq lignes choisies et étudiées précédemment constituent encore le bon choix. En recommençant l'étape de sélection par régression pas à pas, il apparaît que les cinq mêmes lignes semblent encore constituer le choix optimal. Ceci témoigne d'une

certaine stabilité de ce groupe de 5 lignes. Nous n'avons pas testé cette stabilité sur une période encore plus courte, car les effets de la Guerre du Golfe se seraient trop faits sentir sur la prévision.

La relation de régression (en delta-log) calibrée sur les 69 mois de janvier 1988 à décembre 1993 moins les mois d'août, octobre et novembre 1993[10] est :

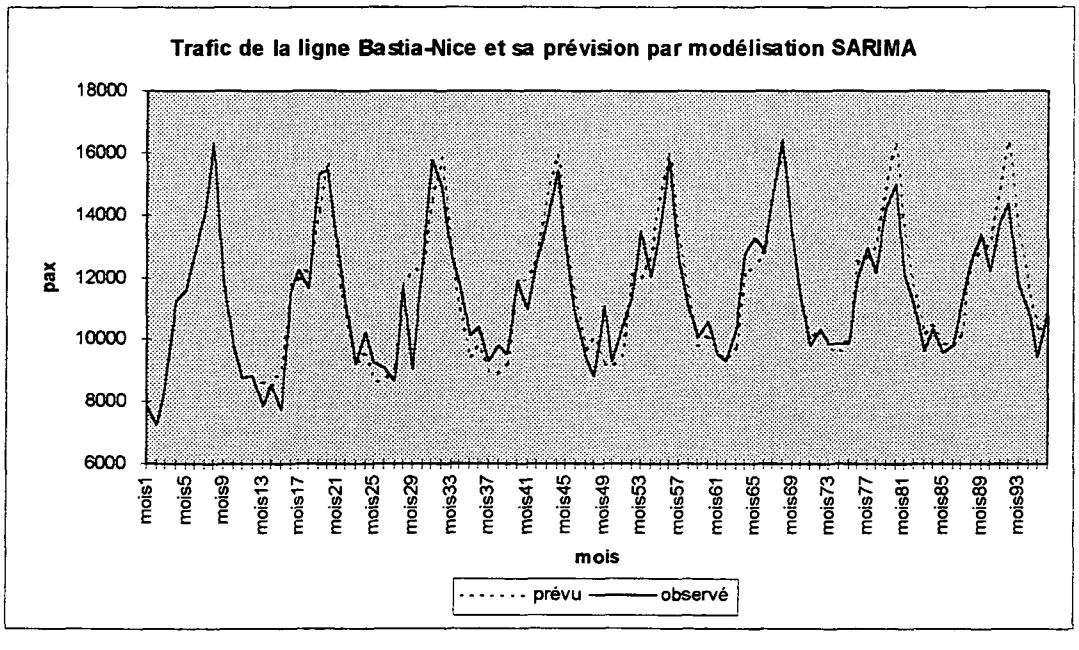
$$\begin{aligned} \text{DOMESTIQUE-8RADCONCUR} = {} & 0.021 + 0.086 \times \text{BASTIA NICE} + 0.304 \times \text{BIARRITZ ORLY} \\ & \quad (7.66)\ (3.79) \qquad\qquad (13.65) \\ & + 0.059 \times \text{BORDEAUX CDG} + 0.190 \times \text{LYON TOULOUSE} \\ & \quad (6.32) \qquad\qquad (7.45) \\ & + 0.136 \times \text{ORLY RENNES}^{11} \\ & \quad (13.87) \end{aligned}$$

Avec : $R^2_{adj} = 0.931$ DW=2.006

Comme pour la série du trafic total, nous avons conservé, pour chaque ligne, les modèles SARIMA déterminés au I, et calculé une prévision pour 1994 en calibrant les modèles sur les sept premières années. On obtient alors une prévision pour le trafic total hors concurrence en utilisant la relation de régression ci dessus.

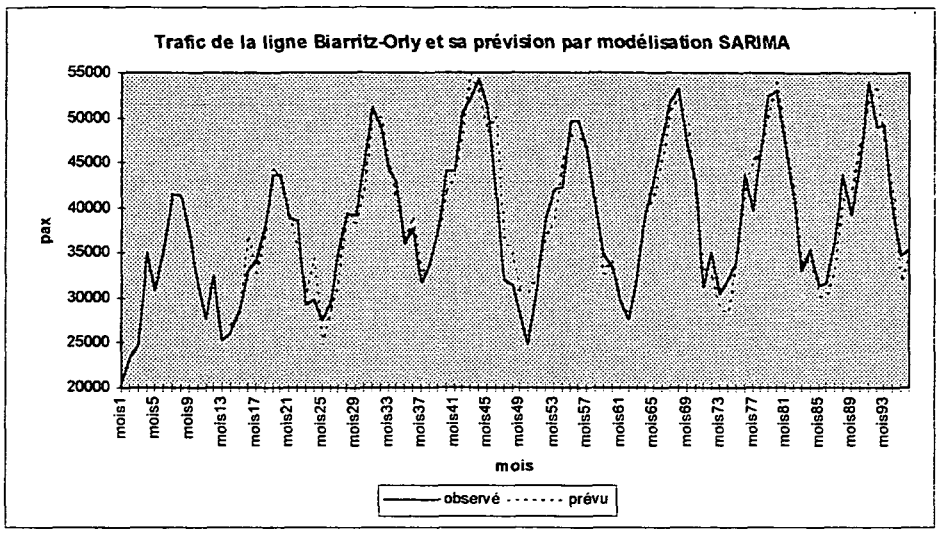
Les prévisions de chaque ligne sont représentées sur la figure 20.

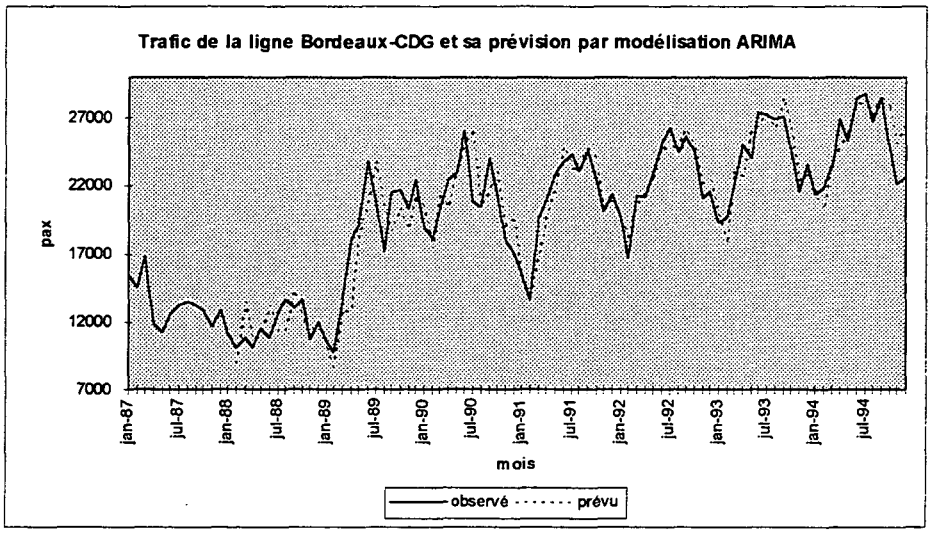
Fig 20 : Prévisions du trafic des cinq lignes

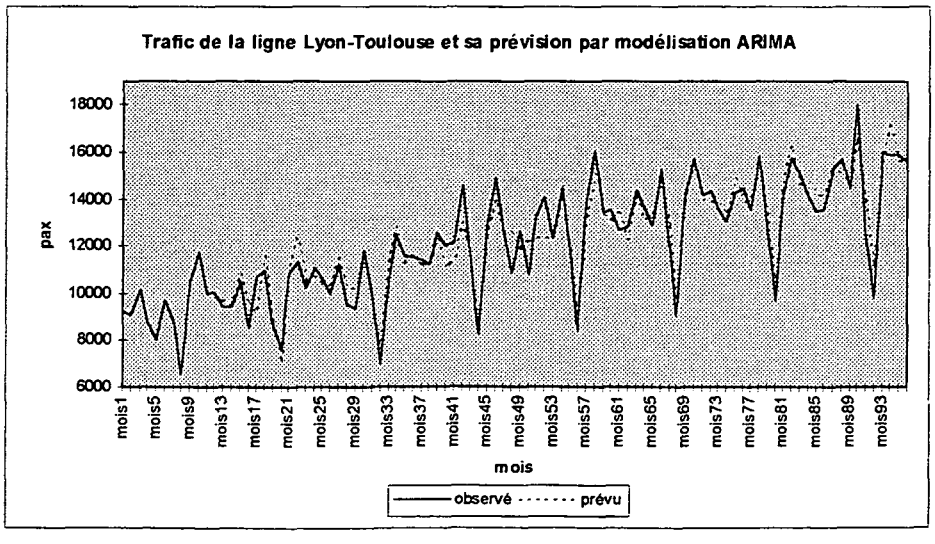


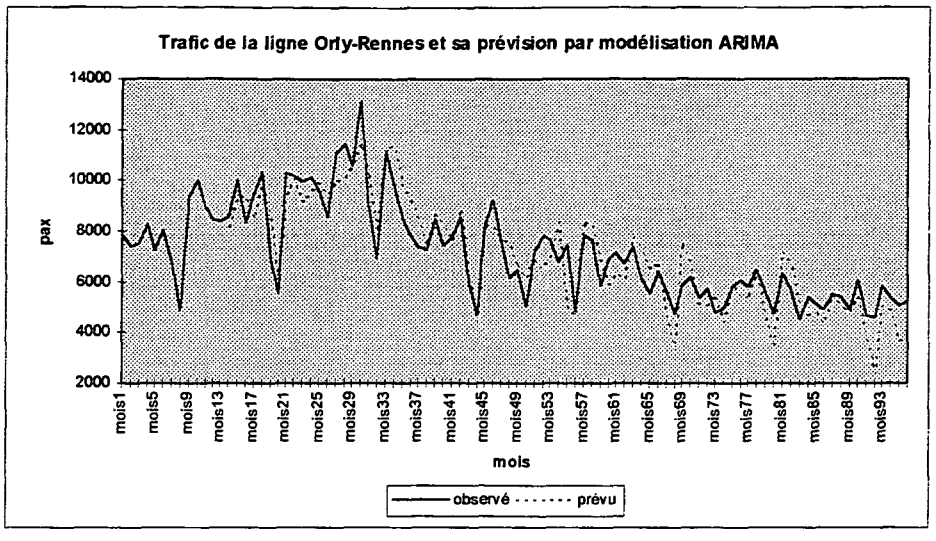
[10] voir la partie II-II-2 (pp 22 et 26) pour l'explication du retrait de ces trois mois

[11] les nombres entre parenthèses sont les t de student associés aux coefficients de la régression (voir p 30)


Trafic de la ligne Biarritz-Orly et sa prévision par modélisation SARIMA
pax
55000
50000
45000
40000
35000
30000
25000
20000
mois1
mois5
mois9
mois13
mois17
mois21
mois25
mois29
mois33
mois37
mois41
mois45
mois49
mois53
mois57
mois61
mois65
mois69
mois73
mois77
mois81
mois85
mois89
mois93
mois
observé
prévu


Trafic de la ligne Bordeaux-CDG et sa prévision par modélisation ARIMA
pax
27000
22000
17000
12000
7000
jan-87
jul-87
jan-88
jul-88
jan-89
jul-89
jan-90
jul-90
jan-91
jul-91
jan-92
jul-92
jan-93
jul-93
jan-94
jul-94
mois
observé
prévu


Trafic de la ligne Lyon-Toulouse et sa prévision par modélisation ARIMA
pax
18000
16000
14000
12000
10000
8000
6000
mois1
mois5
mois9
mois13
mois17
mois21
mois25
mois29
mois33
mois37
mois41
mois45
mois49
mois53
mois57
mois61
mois65
mois69
mois73
mois77
mois81
mois85
mois89
mois93
mois
observé
prévu

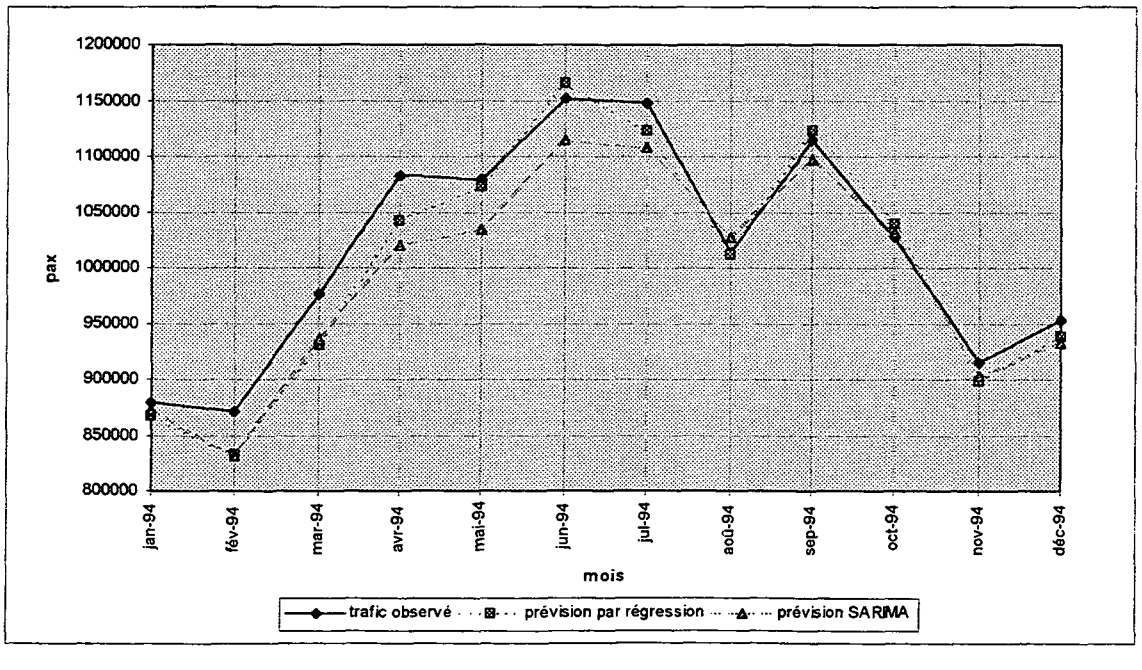


De ces prévisions en nombre de passagers, on tire, à l'aide du trafic *connu* de 1993, des prévisions en delta-log qui permettent, à l'aide de la relation de régression, d'obtenir des prévisions en delta-log pour le trafic total hors concurrence. On transforme alors ces dernières en nombre de passagers, à l'aide encore du trafic *connu* de 1993. Ces prévisions sont représentées sur la figure 21.

II-3- Comparaison des résultats

La figure 21 représente, pour l'année 1994, le trafic total hors concurrence observé, sa prévision par une modélisation ARIMA sur les sept premières années, et sa prévision par régression sur les prévisions des cinq lignes explicatives déterminées à la partie II.

Fig 21 : Comparaison du trafic total observé et de ses prévisions

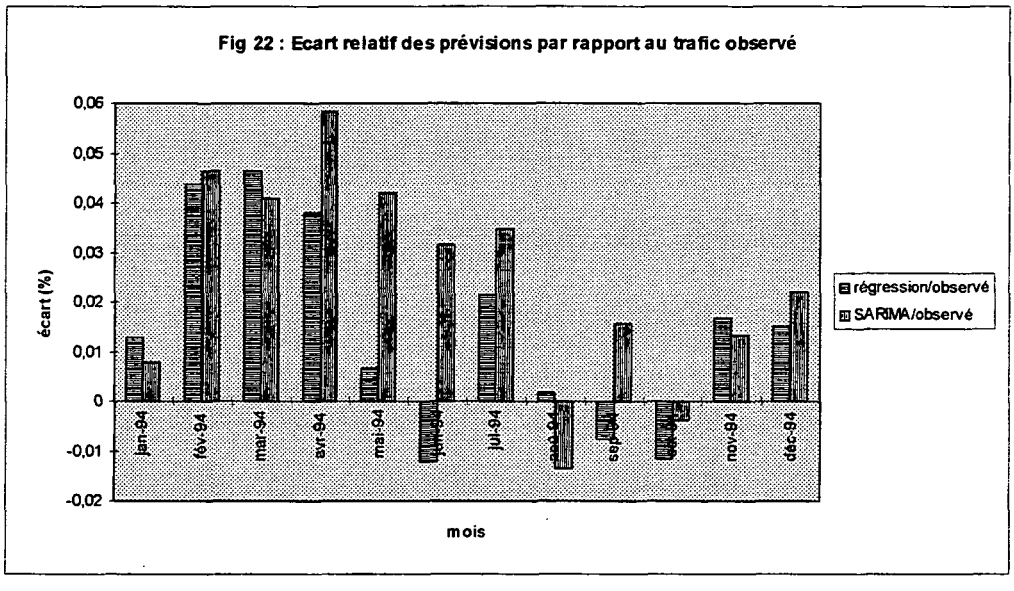


On constate sur la figure 21 que la prévision obtenue grâce à la relation de régression est, pour tous les mois, meilleure que la prévision directe résultant d'une modélisation SARIMA du trafic total.

La figure 22 représente, pour l'année 1994, les écarts relatifs, mois par mois, des prévisions par rapport au trafic observé. Cela confirme que la régression fournit de meilleures prévisions.

L'écart relatif moyen sur un an, est de 1.5% pour la prévision par régression, et de 2.2% pour la prévision directe.

Ces bons chiffres doivent toutefois être tempérés par la constatation que les prévisions du deuxième semestre ont été meilleures que celles du premier semestre. Ainsi, l'écart relatif moyen sur les six premiers mois, tombe, respectivement, à 2.1% et 3.8%.



Nous disposons ainsi d'un outil de prévision à court terme du trafic domestique hors concurrence statistiquement satisfaisant.

III - Etude des huit lignes soumises à la concurrence

Pour compléter cet outil de prévision nous allons nous intéresser à la somme des trafics des huits lignes qui sont ou seront dans un proche avenir affectées par la concurrence entre compagnies aériennes[12].

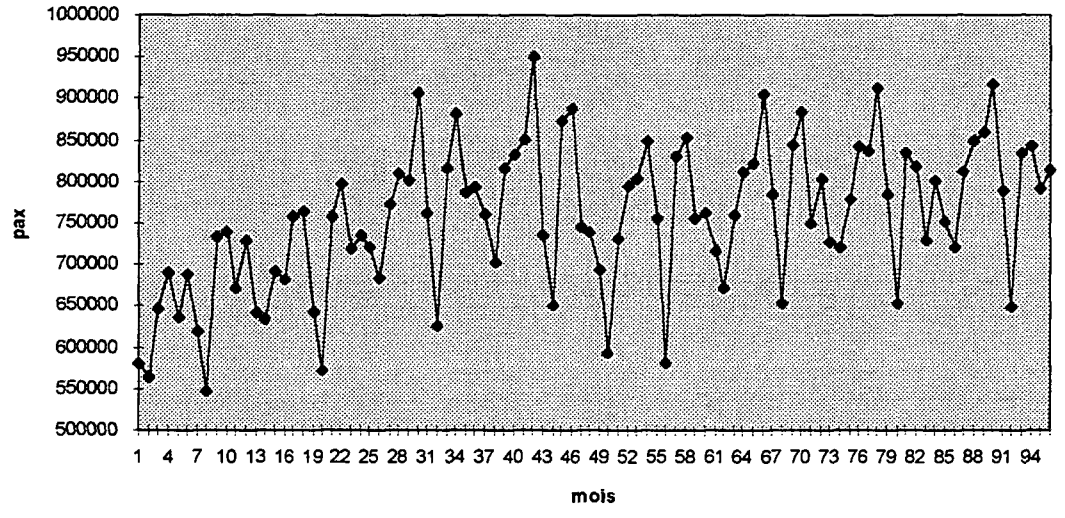
Nous allons modéliser cette somme sur les sept premières années (de janvier 1987 à décembre 1993) par un modèle SARIMA, et prévoir, grâce à ce modèle, le trafic de 1994, afin de le comparer au trafic réellement observé. (Remarquons que nous aurons alors une mesure de l'influence de l'ouverture à la concurrence des lignes TOULOUSE - ORLY et MARSEILLE - ORLY en 1994).

---

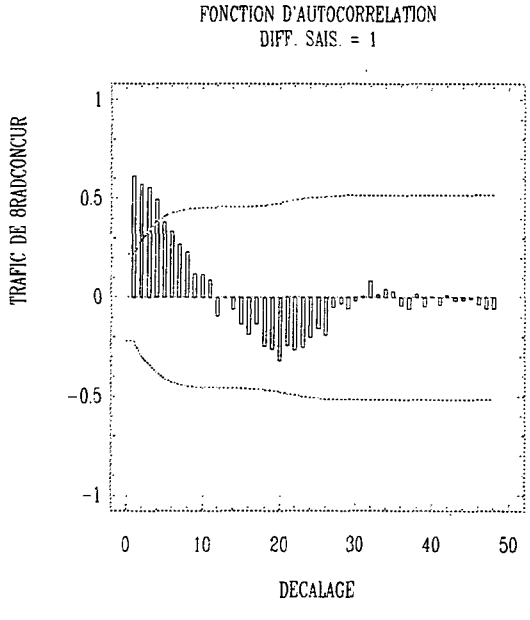
12 voir page 14

La série de trafic à modéliser est représentée sur la figure 23.

Fig 23 : Le trafic de 8RADCONCUR

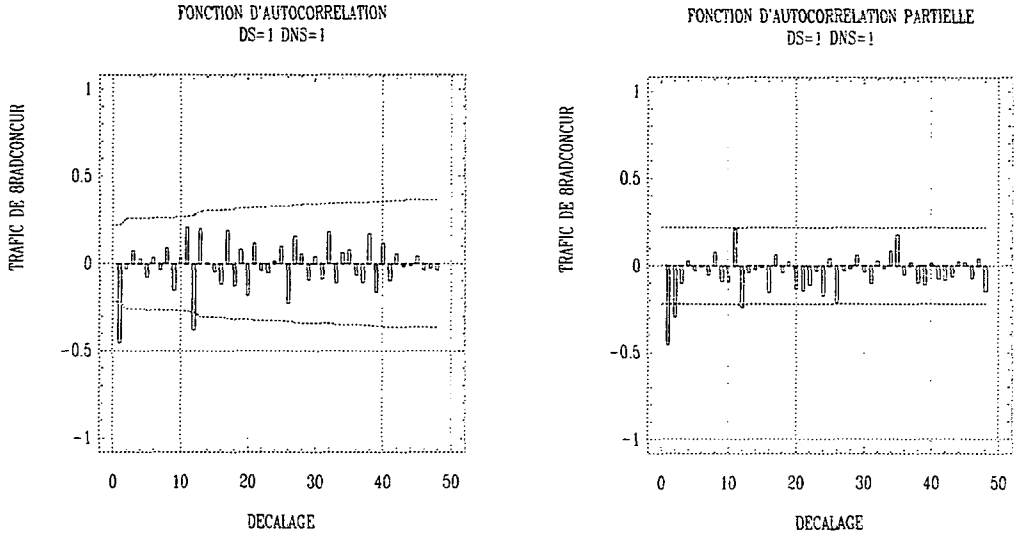


Elle possède une composante saisonnière que nous stabilisons par une différentiation saisonnière d'ordre 1.



Le corrélogramme montre que la série n'est pas stationnaire en tendance. Nous la stabilisons alors par une différentiation régulière d'ordre 1.

Fig 24 : Corrélogramme et corrélogramme partiel de la série du trafic de 8RADCONCUR (différences saisonnière et régulière d'ordre 1)



En non saisonnier, la décroissance des autocorrélations après r_1 est plus rapide que celle des autocorrélations partielles.

C'est la même chose en saisonnier après r_{12}.

Tout ceci nous conduit à essayer le modèle SARIMA $(0,1,1)(0,1,1)_{12}$.

Voilà les résultats des estimations :

Summary of Fitted Model for: 8RADCONCUR

Parameter	Estimate	Stnd.error	T-value	P-value
MA (1)	.55990	.09212	6.07764	.00000
SMA(12)	.46386	.10497	4.41880	.00003
MEAN	-633.87248	1028.08829	-.61655	.53928
CONSTANT	-633.87248			

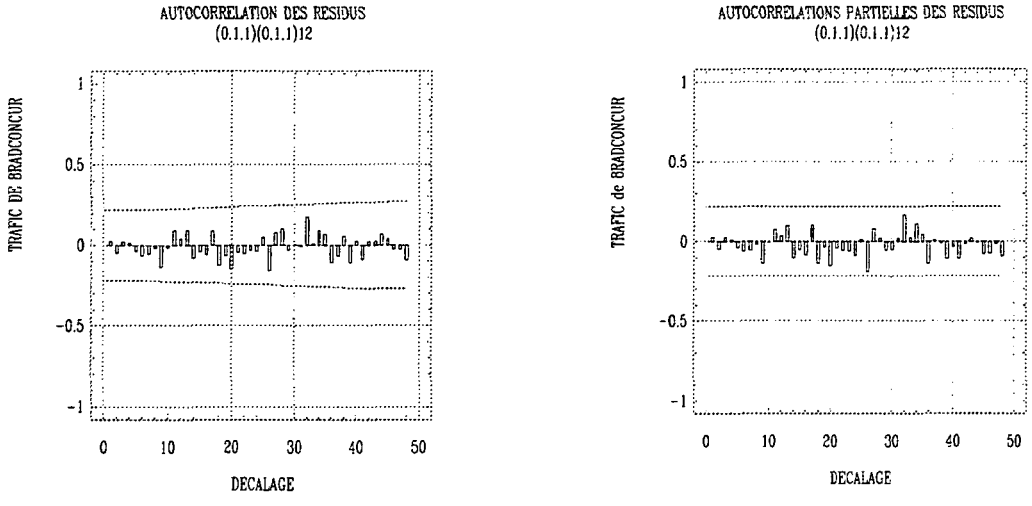
Model fitted to differences of order 1
Model fitted to seasonal differences of order 1 with seasonal length = 12
Estimated white noise variance = 1.15101E9 with 80 degrees of freedom.
Estimated white noise standard deviation (std err) = 33926.6
Chi-square test statistic on first 20 residual autocorrelations = 8.93324
with probability of a larger value given white noise = 0.942333
Backforecasting: no Number of iterations performed: 4

Le modèle s'écrit :

$$(1-B)(1-B^{12})x_t = -633.87 + (1-0.560B)(1-0.464B^{12})\varepsilon_t$$

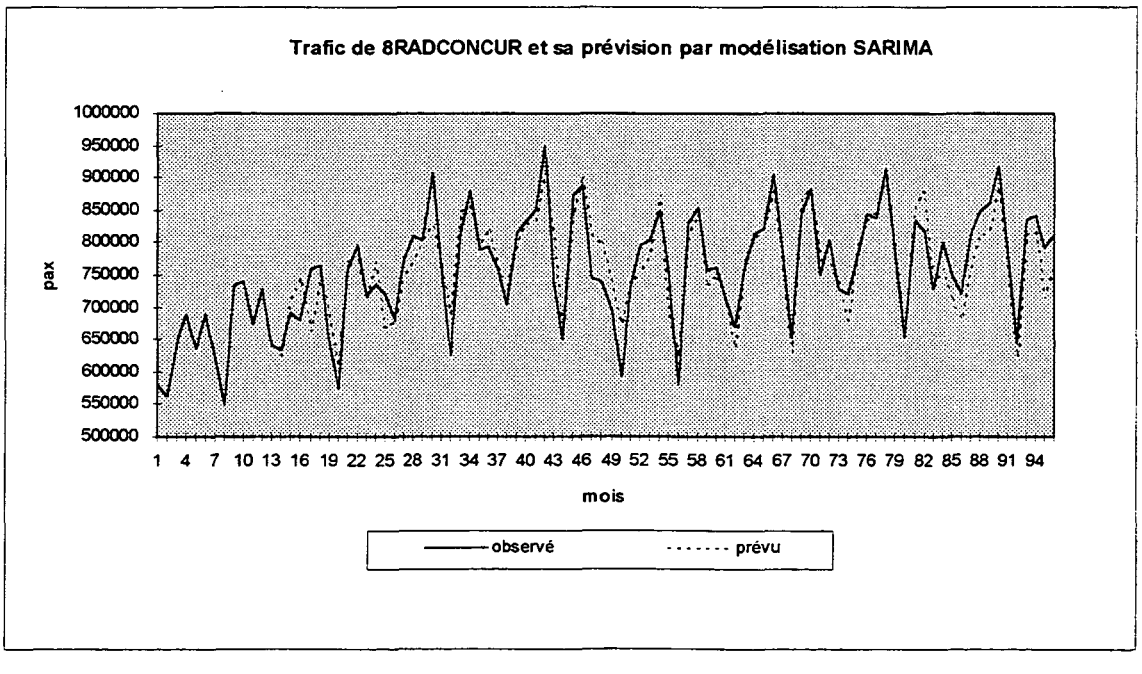
Le portmanteau test montre que l'hypothèse d'un bruit blanc pour les résidus est globalement acceptable (puisque $8.93 < \chi^2_{0.95}(18) = 28.9$). La figure 25 montre qu'il n'y a pas d'autocorrélation significative des résidus.

Fig 25 : Corrélogramme et corrélogramme partiel des résidus

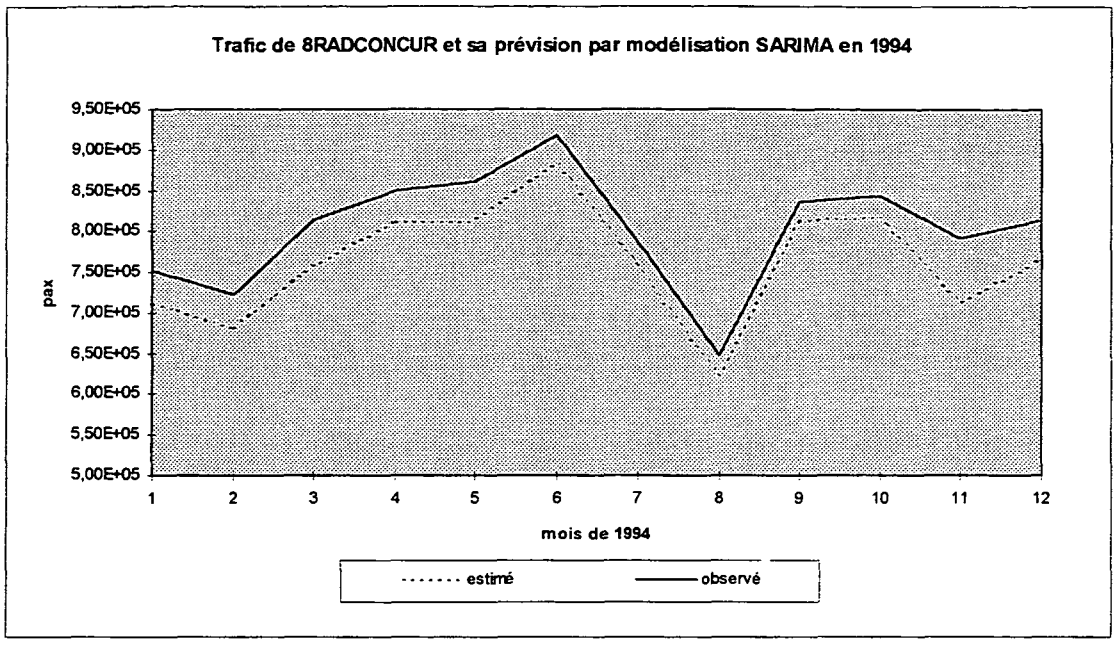


Le modèle est donc acceptable.

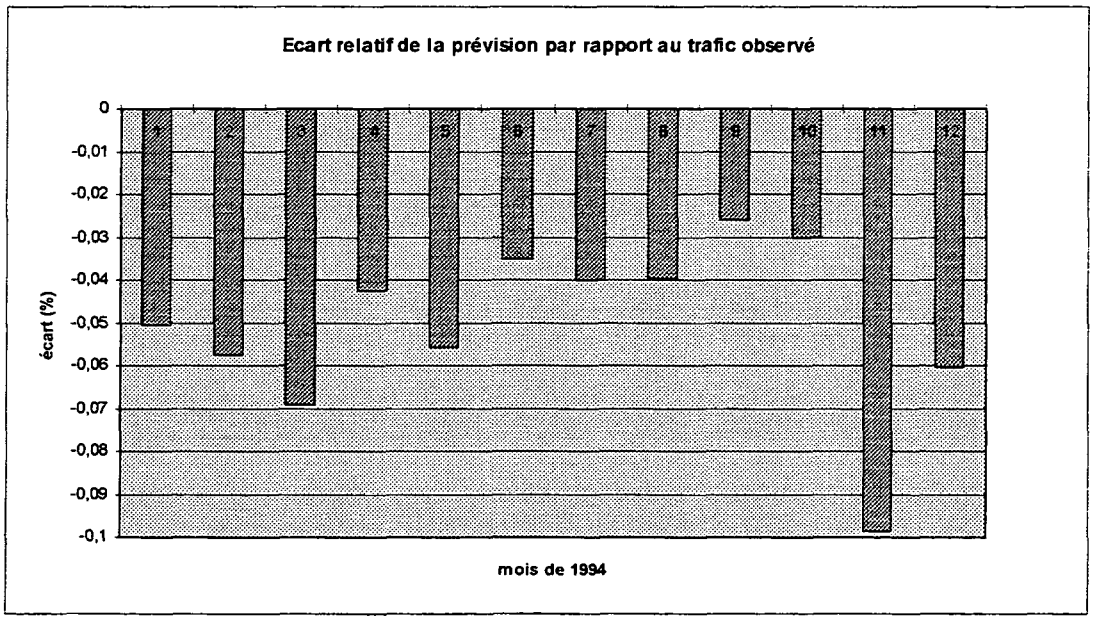
En conservant le modèle SARIMA $(0,1,1)(0,1,1)_{12}$ déterminé ci dessus, nous avons recalibré le modèle sur la période janvier 1987 - décembre 1993, et établi des prévisions pour l'année 1994. Elles sont visualisées ci-dessous :



Jusqu'à décembre 1993, il s'agit de prévision à un mois, faite le mois précédent. En revanche pour l'année 1994, il s'agit de prévisions de 1 à 12 mois, effectuées en décembre 1993. Les prévisions de 1994 ont été agrandies ci-après.



Il apparaît que les prévisions sont toutes inférieures aux observations. La figure suivante représente les écarts relatifs :



La moyenne des écarts relatifs entre prévisions et observations est de 5.04% sur l'année 1994. C'est à comparer aux 2.2% pour le trafic hors concurrence[13]. Ces écarts étant tous dans le même sens (observations supérieures aux prévisions), on peut déduire qu'il s'est produit un changement de tendance, et l'augmentation de trafic par rapport à 1993 a été plus forte pour le groupe partiellement soumis à la concurrence. Ceci peut s'expliquer en partie par l'intensification de la concurrence sur la ligne ORLY NICE, mais aussi par d'autres caractéristiques de ce groupe de huit lignes, qui lui confèrent une plus grande sensibilité aux variables déterminant le trafic.

---

13 voir page 56

CONCLUSION

Nous avons, dans un premier temps, constitué une base de données "nettoyée" à partir d'un certain nombre de critères : de niveau minimum de trafic, une des deux sources complète sur l'ensemble des huit ans, et de niveau d'écart entre les deux sources.

Cette base allégée comporte 60 lignes (à comparer aux 3000 lignes de la base de données fournie par la DGAC), et va servir à la constitution du modèle.

Dans un deuxième temps nous avons estimé un modèle du trafic domestique total à l'exclusion de huit radiales (appelé trafic "total") qui sont déjà, ou seront à plus ou moins court terme, affectées par la concurrence aérienne. Ce modèle relie le trafic total à cinq lignes aériennes et résulte d'une régression pas à pas effectuée sur l'ensemble des 52 lignes de la base allégée non affectée par la concurrence.

Des modèles plus fins, constitués sur des sous-ensembles de lignes homogènes sélectionnées par classification, auraient peut-être permis un meilleur choix des lignes "explicatives" du trafic total. Une tentative de classification ascendante hiérarchique n'a pas aboutie en raison de la présence de données aberrantes, trop nombreuses pour être toutes corrigées dans le cadre de cette étude.

La relation retenue est satisfaisante au plan statistique et peut être utilisée comme outil de suivi du trafic domestique hors concurrence.

Cependant, le choix du modèle retenu (les cinq lignes aériennes et leurs coefficients) est probablement affecté par la présence de nombreuses valeurs aberrantes dont seulement quelques unes ont été corrigées dans le cadre de cette étude, faute de temps. Il serait intéressant de procéder à un traitement plus systématique des valeurs atypiques, ligne par ligne. En revanche, l'existence de deux sources différentes, et divergentes, pour le même trafic mensuel, pose un problème qui n'admet pas de solution technique; sa résolution ne peut s'envisager qu'en amont de l'étude statistique.

Nous avons enfin proposé et comparé deux méthodes de prévision à court terme du trafic total hors concurrence. La première méthode consiste à modéliser directement le trafic "total" par un modèle SARIMA. La seconde consiste à appliquer la relation de régression, établie précédemment, à des prévisions de trafic des cinq lignes aériennes explicatives, obtenues individuellement par des modélisations SARIMA.

Nous avons, pour les évaluer, appliqué ces méthodes à la prévision du trafic mensuel de 1994, en calibrant les modèles sur les trafics de janvier 1987 à décembre 1993.

Il s'est avéré que la méthode utilisant la régression donne, sur cet exemple, de meilleurs résultats. Même si cela n'a pas valeur de preuve, c'est une incitation à privilégier cette méthode.

Nous disposons ainsi d'un outil de suivi et de prévision à court terme du trafic domestique hors concurrence. Les étapes qui y ont conduit ont été détaillées, ce qui devrait faciliter le recalibrage des modèles

Cette étude se termine par une étude du trafic domestique global affecté par la concurrence aérienne à compter de 1993: une étude détaillée des huit lignes concernées n'a pas été réalisée faute de temps. Le modèle retenu peut être utilisé pour appréhender les effets qui se manifestent sur cet ensemble de huit lignes: effets de la concurrence (limités à ORLY NICE à compter de 1993 et à ORLY TOULOUSE et ORLY MARSEILLE à compter de 1995) d'une part, mais aussi effets des modifications des variables déterminantes de l'offre et de la demande de transport aérien (tarifs, fréquences, conjoncture économique...).

Logiciels informatiques utilisés

- ✓ Le rapport a été rédigé à l'aide de Lotus Ami Pro 3.0 et de Microsoft Excel 5.0
- ✓ La classification hiérarchique ascendante a été réalisée grâce à la procédure CLUSTER du logiciel SAS.
- ✓ Les régressions pas à pas et multiples, les calculs de corrélation et les modélisations SARIMA ont été réalisées à l'aide de STATGRAPHICS 7.1

Bibliographie élémentaire

- ✓ Christian Labrousse, *Introduction à l'économétrie*, DUNOD 1985
- ✓ T.H. Wonnacott et R.J. Wonnacott, *Statistique*, ECONOMICA 1991
- ✓ Denis Bosq et Jean-Pierre Lecoutre, *Analyse et prévision des séries chronologiques*, MASSON 1992
- ✓ Régis Bourbonnais, *Econométrie*, DUNOD 1993
- ✓ Michel Tenenhaus, *Méthodes Statistiques En Gestion*, DUNOD 1994