REVUE DE STATISTIQUE APPLIQUÉE

PATRICK LE BRETON

Un exemple d'application de l'analyse des données dans les sondages : évaluation du trafic sur les routes secondaires dépourvues de comptages permanents

Revue de statistique appliquée, tome 33, n° 3 (1985), p. 5-14

<http://www.numdam.org/item?id=RSA_1985__33_3_5_0>

UN EXEMPLE D'APPLICATION DE L'ANALYSE DES DONNÉES DANS LES SONDAGES : ÉVALUATION DU TRAFIC SUR LES ROUTES SECONDAIRES DÉPOURVUES DE COMPTAGES PERMANENTS

Patrick LE BRETON

S.E.T.R.A., (Service d'Etudes Techniques des Routes et Autoroutes)

RÉSUMÉ

Le présent article montre l'intérêt de l'utilisation de techniques d'analyse des données et en particulier de la régression par boule dans le domaine des sondages lorsque l'on dispose d'un grand nombre de variables correlées avec la variable à estimer. L'exemple traité consiste à évaluer par sondage le trafic annuel sur des sections de route à trafic modéré en s'appuyant sur la connaissance du trafic sur les routes disposant de comptages permanents. Un programme de simulation est mis en œuvre pour évaluer la précision des résultats.

ABSTRACT

This article shows the advantage of using data analysis methods, particulary the regression by neighbours in the field of sampling when a great number of variables correlated to the one to estimate are available. The exemple treated here consists in the evaluation sampling of the annual traffic volume on road sections with moderate traffic by relying on the knowledge of the traffic on road who dispose of permanent counting. A simulation program is set to evaluate the precision of the results.

1. INTRODUCTION

L'intérêt de l'utilisation d'informations supplémentaires dans la théorie des sondages sous forme d'estimation par le quotient ou par la régression n'est plus à démontrer. Lorsque l'on est en possession d'un grand nombre de variables, l'utilisation de techniques d'« Analyse des données » et en particulier de la « régression par boule » permet dans le même ordre d'idée de construire un système de sondage cohérent et particulièrement efficace.

I. POSITION DU PROBLÈME

Le Ministère des Transports a besoin pour de nombreuses applications (entretien des chaussées, étude de la fréquence des accidents, etc.) d'avoir une évaluation du trafic circulant sur les différentes routes. Le réseau de routes à grande circulation est équipé de compteurs permanents qui fournissent le débit de chaque jour de l'année. En raison du coût que cela entraînerait il n'est pas possible de disposer en grand nombre de tels compteurs sur le réseau secondaire beaucoup plus vaste; des sondages doivent donc être réalisés sur la plus grande partie de ce réseau. Le but de

Mots clés : **Sondage, Analyse des données, Régression par boule.**

la présente étude est de proposer des méthodes de comptage appropriées pour l'évaluation du trafic journalier en moyenne annuelle noté par la suite T.J.M.A.

Après avoir effectué l'« Analyse des données » sur les postes des compteurs permanents nous serons amenés à proposer deux méthodes de comptage. La première résulte de la sélection de la période de l'année la plus apte à reconstituer le trafic annuel. La seconde s'inspire à la fois de techniques classiques de sondages concernant la stratification et l'utilisation d'informations supplémentaires (estimation par le quotient ou par la régression) et en prolongement de la « régression par boule ». Celle-ci permet de trouver automatiquement des meilleures informations sur lesquelles le sondage peut s'appuyer.

II. ANALYSE DES POSTES DE COMPTAGE PERMANENT

Pour l'étude on considère de 263 sections du réseau principal ou secondaire notées S pour lesquelles on connaît les trafics journaliers. Dans cet ensemble se trouvent 29 sections du réseau secondaire (T.J.M.A. inférieur à 2 000 véhicules/jour). L'analyse est réalisée pour partie sur les années 1970 à 1975 et 1980; une autre partie n'est réalisée que pour l'année 1971 mais est par son caractère général utilisable pour toute autre année.

La connaissance du trafic sur les 263 sections est analysée au moyen du tableau de contingence t (j, s) avec

j numéro du jour de l'année ($1 \leqslant j \leqslant 365$)
s numéro de la section de comptage permanent ($1 \leqslant s \leqslant 263$)
t (j, s) est alors le trafic (en nombre de véhicules) observé le jour j sur la section s.

Le T.J.M.A. (trafic journalier moyen sur l'année) est à une constante multiplicative près égal à l'une des marges de ce tableau :

$$t(s) = \left\{ \sum_j t(j, s)/j = 1{,}365 \right\}$$

L'autre marge est constituée par le profil moyen du trafic sur l'ensemble des routes de S.

Une analyse factorielle des correspondances est effectuée sur le tableau t (j, s).

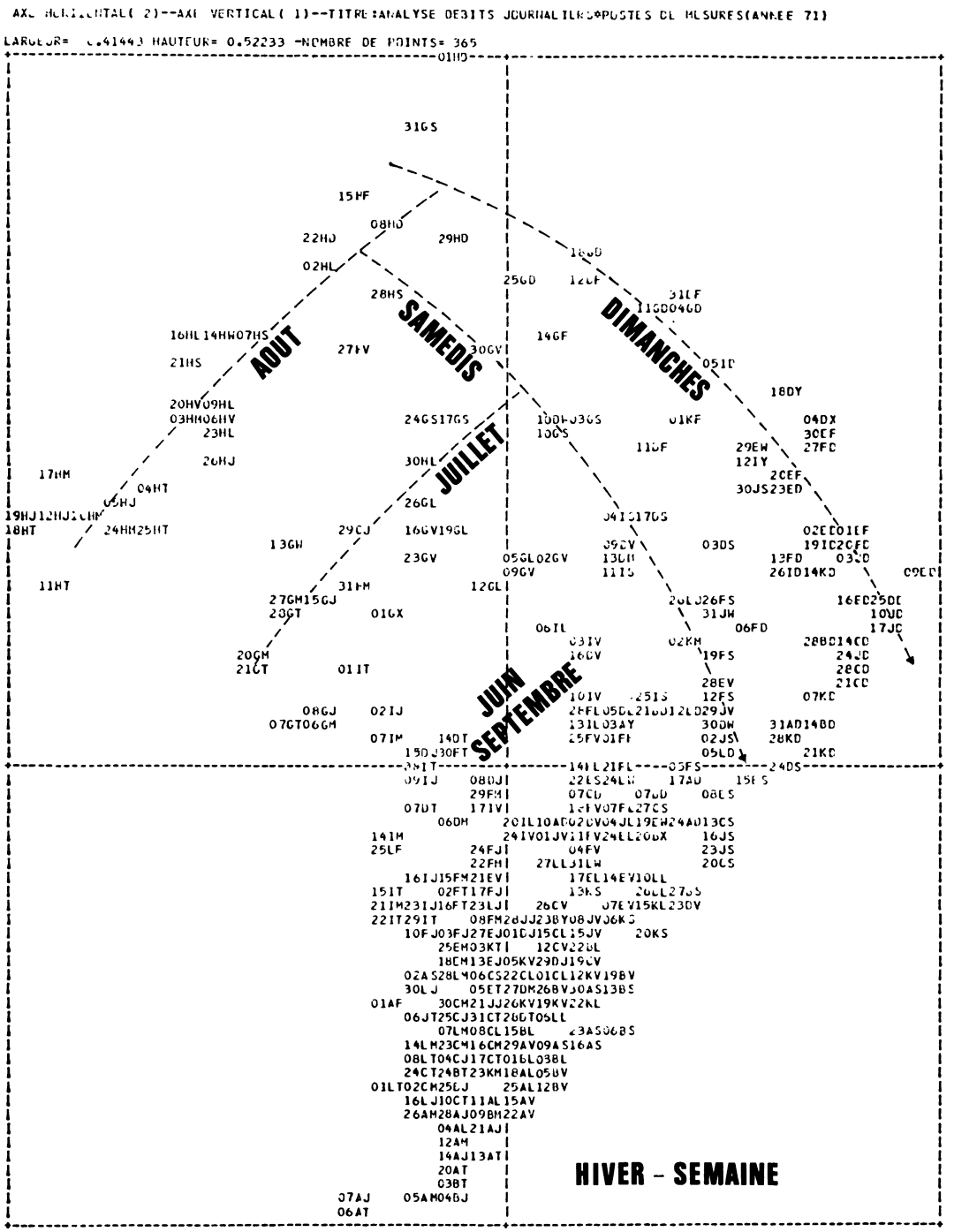
Les axes dégagent dans l'ordre décroissant d'importance les grandes tendances.

Le 1er axe* (36 % de l'inertie totale, valeur propre égale à 0,0125) oppose les jours de semaine d'hiver aux jours d'Eté de week-end et de départ en vacances. Il apparaît comme très corrélé avec le trafic journalier, puisque l'hiver correspond à des trafics journaliers faibles et à l'inverse l'été connaît des trafics journaliers élevés.

Le 2^{e} axe* (16 % de l'inertie totale, valeur propre égale à 0,0055) oppose les jours de semaine d'été aux jours de week-end excluant l'été.

Un examen du plan déterminé par les deux premiers axes fait apparaître des lignes de force : Août, Juillet, Dimanche, Samedi (voir graphique ci-après). Ainsi les dimanches de Juillet sont pratiquement à l'intersection des deux lignes de force correspndantes.

(*) Pour la codification du graphique des deux premiers axes factoriels, voir fin du § II.


AXE HORIZONTAL(2)--AXE VERTICAL(1)--TITRE:ANALYSE DES ITS JOURNALIERS POSTES DE MESURES(ANNEE 71)
LARGEUR= 0.41443 HAUTEUR= 0.52233 -NOMBRE DE POINTS= 365
AOUT
SAMEDIS
DIMANCHES
JUILLET
JUIN
SEPTEMBRE
HIVER - SEMAINE

Le centre de gravité n'étant autre que le T.J.M.A., les jours qui en sont proches reproduisent le T.J.M.A. de façon satisfaisante à un coefficient de proportionnalité près. Ces jours corrrespoondent aux jours en semaine des mois de Septembre et Juin.

Cette propriété s'explique car les mois de Juin et de Septembre sont intermédiaires entre les mois de vacances d'été proprement dit et les autres. On a donc ici une première possibilité d'évaluer le T.J.M.A.

CODIFICATION DU GRAPHIQUE

Quantième, mois, jour de la semaine

Mois :	A janvier	Jour de la semaine :	L lundi
	B février		M mardi
	C mars		T mercredi
	D avril		J jeudi
	E mai		V vendredi
	F juin		S samedi
	G juillet		D dimanche
	H août		X départ congés scolaires
	I septembre		Zone Paris
	J octobre		Y retour congés scolaires
	K novembre		Zone Paris
	L décembre		W veille jours fériés
			F jours fériés

Ainsi 07GT désigne le mercredi 7 juillet 14IM le mardi 14 septembre.

III. PRÉPARATION DU SONDAGE PAR LA STRATIFICATION DES JOURS DE L'ANNÉE

La stratification a pour but de recouvrir les principales tendances de l'analyse factorielle.

L'utilisation de la méthode des nuées dynamiques à partir d'étalons représentatifs et de l'analyse factorielle ci-dessus suggère les strates suivantes :

— jours d'hiver (1er janvier au 24 mars);
— jours d'été (15 juillet au 31 août);
— dimanches et jours fériés;
— autres jours ou jours moyens.

Compte tenu du temps de pose et de dépose des compteurs les dimanches et jours fériés seront inclus dans la même période que les jours de semaines c'est-à-dire que les 4 strates deviennent 3 Périodes de 6 jours chacune incluant obligatoirement le Dimanche. On obtient ainsi un dimanche d'hiver, un dimanche d'été, un dimanche moyen. Le jour de manipulation des compteurs est le Mercredi.

La stratification des jours a été rendue nécessaire car le sondage sera effectué avant toute exploitation statistique. A l'inverse la stratification des sections n'est pas nécessaire. On aurait pu également décomposer les sections en classes et affecter par la suite à la section qui a fait l'objet d'un sondage une de ces classes, mais cette affectation a toujours une part d'arbitraire et peut être sujette à caution. L'utilisation de la « régression par boule », une fois recueillies les données, rend cette démarche inutile :

à une section sondée ne sera pas affectée une classe de sections mais un certain nombre de sections de même typologie. L'idée résultant de la formule de NEYMAN (référence 6 page 145) conduit à adopter un taux de sondage plus élevé pour les strates « Eté » et « dimanche et jours fériés » qui présentent une forte variabilité. Ainsi les jours d'été sont sondés autant que les jours moyens pourtant beaucoup plus nombreux. La période d'été, étant la plus courte, détermine le nombre de sondages par compteur. Il s'élève à 6.

IV. EXPLOITATION STATISTIQUE DES SONDAGES SUR PLUSIEURS STRATES

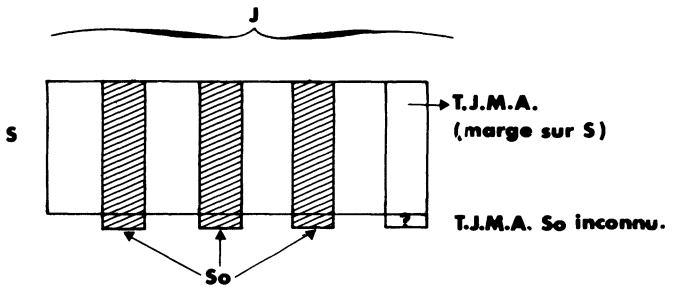
IV.1. Position du problème

Une section sondée n'est, à la différence des sections des routes à grande circulation servant de sections de base, connue que sur un nombre restreint de jours. Elle n'est donc comparable à des sections de base que sur leur zone d'observation commune. Or l'analyse factorielle effectuée plus haut prend en considération tous les jours de l'année.

Désignons par :

S l'ensemble des sections du réseau de comptage permanent,
J l'ensemble des jours de l'année,
$J_0(s_0)$ l'ensemble des jours sondés pour une section sondée s_0 n'appartenant pas à S.

Le problème peut être schématisé de la façon suivante :



plusieurs types de méthodes sont utilisables.

1) Une analyse factorielle sur $J_0(s_0) \times S$ avec s_0 en ligne supplémentaire et les T.J.M.A. en colonne supplémentaire. Le T.J.M.A. de s_0 peut alors être reconstituée par la méthode de « l'estimation d'une donnée manquante à l'intersection d'une ligne et d'une colonne supplémentaire » (réf. 1, III.7).

2) Une régression par boule à partir du tableau $J_0(s_0) \times S$.

Ces deux premières méthodes nécessitant de faire une analyse factorielle pour chaque $J_0(s_0)$, l'exploitation statistique et les coûts risquent d'être très lourds. Nous proposons donc de garder comme tableau de référence le seul tableau $J \times S$.

Pour chaque $J_0(s_0)$ on calcule en utilisant la formule de transition les coordonnées induites par $J_0(s_0)$ des sections de S et les coordonnées de la section s_0 ayant eu le sondage $J_0(s_0)$:

$$G'_\alpha(s) = \left\{ \sum \frac{1}{\sqrt{\lambda\alpha}} \frac{t(j_0, s)}{t^*(., s)} F_\alpha(j_0) / j_0 \in J_0(s_0) \right\}$$

avec :

$$t^*(., s) = \{\Sigma\, t(j_0, s) / j_0 \in J_0(s_0)\} .$$

$G'_\alpha(s)$ coordonnée approchée de s; $F_\alpha(j_0)$ coordonnée de j_0 sur l'axe factoriel α issu de l'analyse factorielle J × S. Avec cette formule (approchée par rapport aux méthodes précédentes) les sections de S et s_0 sont directement comparables par la distance euclidienne calculées à partir des différents axes de l'analyse factorielle. On aura la distance approchée suivante entre s et s_0 :

$$d^2(s, s_0) = \sum_\alpha (G'_\alpha(s) - G'_\alpha(s_0))^2$$

d'où la possibilité d'obtenir les voisins.

On ne peut comparer deux sections que sur leurs jours communs d'observation. En effet rien n'oblige qu'une section placée à partir de deux groupes de jours différents se retrouve au même endroit sur les graphiques.

On aurait pu également obtenir les sections voisines en utilisant la distance du X^2 sans recours aux axes factoriels. Le résultat n'aurait pas été meilleur car on prenait alors en considération les phénomènes aléatoires ou mal appréhendés qui correspondent aux derniers axes de l'analyse factorielle. De plus la recherche des sections voisines aurait été plus lourde.

IV.2. Evaluation du T.J.M.A. de la section sondée

Si la distance entre deux sections s et s_o était exacte et si elle était nulle cela veut dire d'après le principe de base de l'analyse factorielle que leurs trafics journaliers seraient proportionnels et donc que leurs T.J.M.A. seraient dans le même rapport, ce qui permet de déduire l'un à partir de la connaissance de l'autre.

Si le sondage est suffisamment représentatif on peut espérer que si les distances approchées sont nulles, les T.J.M.A. vérifient toujours la propriété précédente.

Nous estimerons donc pour chaque jour sondé j_0 et pour chaque section voisine v le T.J.M.A. de la section s_0 par :

$$\text{T.J.M.A.}(v_0, j_0) = \text{T.J.M.A.}(v) \times \frac{t(j_0, s_0)}{t(j_0, v)}$$

Cette valeur sera pondérée par l'inverse de l'écart-type $\sigma(j_0)$ des T.J.M.A. estimés à partir des sections voisines pour le jour j_0. En effet plus cet écart-type est faible plus le jour correspondant est apte à reproduire le T.J.M.A. Si n_v représente le nombre de sections voisines et n_{j0} le nombre de jours sondés on aura donc pour chaque section sondée $n_{j0} \times n_V$ estimations de son T.J.M.A. L'étude de la dispersion de ces valeurs permet d'avoir une idée de la précision.

Le T.J.M.A. sera estimé en moyenne pour la section s_0 par :

$$\frac{1}{n_v} \left(\sum_{j \in J_0(s_0)} \frac{1}{\sigma(j_0)} \sum_v \text{T.J.M.A.}(v_0, j_0) \right) \Big/ \sum_{j \in J_0(s_0)} 1/\sigma(j_0)$$

Le biais correspondant à une telle estimation peut être calculé selon HARTHLEY et ROSS (référence 5 page 174) et par suite le redressement correspondant peut être effectué.

V. COMPTAGE SUR UNE SEULE PÉRIODE

Lorsque l'on se fixe une seule période, l'analyse factorielle a montré que les mois à considérer a priori sont ceux de Juin et Septembre. Le problème de comptage se résoud ensuite en termes de moindres carrés.

L'étude sera faite ici sur les années 70 à 75 et 80. Les analyses factorielles faites sur plusieurs années montrent que le vendredi de septembre compris selon l'année entre le 17 et le 23 est le plus proche du centre de gravité. C'est donc le jour le plus apte à reconstituer le T.J.M.A. En effet, une régression effectuée sur ce jour de trafic T_j et sur l'ensemble des sections de comptage permanents des 7 années donne la formule suivante T.J.M.A. = 0,907 T_j + 318 avec un coefficient de corrélation multiple égal à 0,987. Comme le T.J.M.A. moyen est de 11 000 véhicules/jour nous en déduisons que l'impact du terme constant est négligeable.

Pour la suite, nous considérons une période de 7 jours, qui permet de lisser l'effet du jour de la semaine, en « moyenne mobile ». Comme nous nous intéressons aux routes à faible trafic nous ne chercherons pas à minimiser l'erreur absolue mais plutôt l'erreur relative.

Soit donc à minimiser

$$\varepsilon_j^2 = \sum_s \left(\frac{\text{T.J.M.A. (s)} - a_j\, T_j(s)}{\text{T.J.M.A. (s)}} \right)^2 / N_s$$

avec

T.J.M.A. (s) : T.J.M.A. de la section s
$T_j(s)$: trafic moyen sur 7 jours en moyenne mobile d'ordre 7 autour du jour j, a_j : coefficient de redressement à évaluer N_s le nombre de sections (N_s = 263)

après dérivation nous obtenons :

$$a_j = \frac{\sum_s T_j(s) \,/\, \text{T.J.M.A.(s)}}{\sum_s (T_j(s)/\text{T.J.M.A.(s)})^2}$$

$$\varepsilon_j = \sqrt{1 - a_j \frac{\sum_s t_j(s) \,/\, \text{T.J.M.A.(s)}}{N_s}}$$

Ces régressions sont faites année par année. A noter que a_j est faible si $T_j(s)$ est en moyenne plus élevé que le T.J.M.A. correspondant. C'est le cas pour la période d'été. Le graphique ci-après donne pour les 7 années considérées en haut les valeurs des a_j en bas les valeurs des ε_j multipliés par 1,96 pour obtenir la précision calculée sur l'ensemble des sections de comptage permanent.

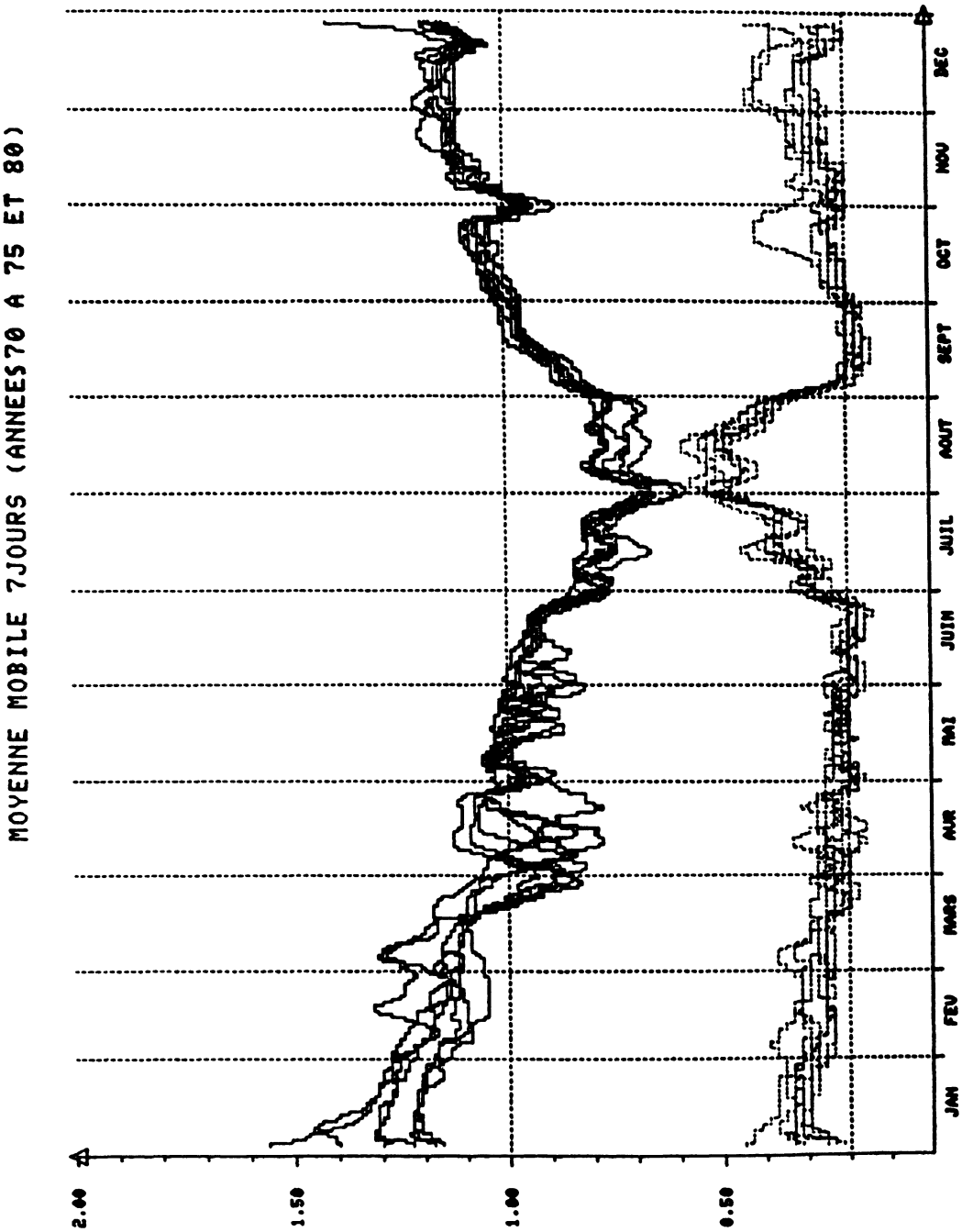
On constate que le mois de Septembre possède trois propriétés importantes :

1) Les a_j sont permanents d'une année sur l'autre.

2) C'est le mois où la précision est la meilleure. Cette propriété était prévisible à la suite de l'analyse factorielle. Pour les 29 sections à faible trafic (T.J.M.A. inférieur à 2 000 véhicules/jour). La valeur de 1,96 ε_j en moyenne sur le mois de Septembre est égale à 0,25.

3) Si maintenant nous cherchons à estimer a_j pour le mois de Septembre en fonction du numéro du jour dans le mois n (1er Septembre : n = 1) nous obtenons :

a_j = 0,00612 n + 0,84 avec un coefficient de corrélation multiple égal à 0,82.


MOYENNE MOBILE 7JOURS (ANNEES70 A 75 ET 80)
2.00
1.50
1.00
0.50
JAN
FEV
MARS
AVR
MAI
JUIN
JUIL
AOUT
SEPT
OCT
NOV
DEC

Cette régression est réalisée sur 7 fois 30 observations. La formule exprime la baisse régulière du trafic pendant le mois de Septembre.

VI. RÉSULTATS : TEST DES MÉTHODES PROPOSÉES

Le recours à la simulation est rendu nécessaire car il y a plus de 2 000 sondages différents possibles, le choix à l'intérieur des strates étant laissé libre sauf, nous l'avons vu pour la strate « dimanche et jours fériés ».

Les tests sont effectués par simulation sur les 29 sections de comptage permanent qui sont également à faible trafic.

Etant donné une section à simuler, les sections voisines sont choisies parmi les 262 autres sections.

De façon à mettre en évidence, l'intérêt des méthodes proposées nous l'avons comparée à la stratification classique. Nous avons également fait varier le nombre de voisins.

n_k tirages aléatoires de 3 × 6 jours chacun étant réalisés par section, l'erreur quadratique moyenne est donnée par la formule :

$$\varepsilon_s = \sqrt{\sum_k (T_k, s - \text{T.J.M.A.}(s))^2/(n_k - 1)}$$

ou T_k, s désigne le T.J.M.A. estimé par le $k^{ième}$ tirage pour la section s.

La moyenne de l'erreur sur l'ensemble des sections est donnée par :

$$\varepsilon = \sum_s \varepsilon_s / n_s$$

La précision relative prise égale à 1,96 fois l'erreur relative correspondant à ε_s est également calculée. n_k est pris égal à 5 ce qui fait 5 × 29 simulations.

La simulation montre qu'il est intéressant de considérer pour calculer la distance entre sections 4 facteurs. Les résultats sont pratiquement les mêmes lorsque l'on enlève le terme $(1/\sqrt{\lambda\alpha})$ dans la formule de transition.

Les résultats sont les suivants :

		erreur moyenne	précision (2 fois l'erreur relative)
3 fois 6 jours cf. § III	stratification usuelle	180	26 %
	régression par boule un voisin cf. § IV.2	131	19 %
	régression par boule 20 voisins cf. § IV.2	100	14 %
7 Jours septembre avec formule cf. § V		147	21 %

Un gain de précision peut encore être escompté si l'on augmente le nombre de sections de base ou si elles sont prises dans un ensemble de sections à faible trafic.

La précision des résultats pour 7 jours de septembre est légèrement différente de celle obtenue au paragraphe V du fait que la formule de calcul de l'erreur n'est pas la même. En effet la régression donne naturellement une erreur quadratique alors que nous cherchons une erreur moyenne par section sondée. Si on cumule les deux méthodes nous obtenons dix sondages possibles par compteur.

CONCLUSION

Les résultats montrent l'intérêt des deux méthodes résultant directement de l'analyse des données à savoir choix d'une période privilégiée de l'année : le mois de septembre ou utilisation de la régression par boule.

RÉFÉRENCE

[1] Ch. BASTIN, J.P. BENZECRI, Ch. BOURGARIT, P. CAZES (1980). — *Pratique de l'analyse des données, Abrégé théorique. Etude de cas. Modèle.* Vol. 2. (III n° 7), Dunod.

[2] M.O. LEBEAUX (1977). — Notice sur l'utilisation du programme POUBEL, *C.A.D.*, Vol. II, n° 4, pp. 467-481.

[3] P. CAZES (1975). — Régression par boule et par l'analyse des correspondances, R.S.A., vol. XXIV n° 4, pp. 5-22.

[4] L. LEBART, A. MORINEAU, J.P. FENELON (1979). — *Traitement des données statistiques,* Dunod.

[5] G. COHRAN (1977). — *Sampling techniques.* 3e édition.

[6] J. DESABIE (1971). — *Théorie et pratique des sondages,* Dunod.