REVUE DE STATISTIQUE APPLIQUÉE

M. A. CAMBOIS
H. FONTAINE

# Présentation d'une méthode de segmentation effectuée à partir de deux fichiers différents. Application à la sécurité routière

*Revue de statistique appliquée*, tome 28, n° 4 (1980), p. 37-49
<http://www.numdam.org/item?id=RSA_1980__28_4_37_0>

© Société française de statistique, 1980, tous droits réservés.

L'accès aux archives de la revue « Revue de statistique appliquée » (http://www.sfds.asso.fr/publicat/rsa.htm) implique l'accord avec les conditions générales d'utilisation (http://www.numdam.org/conditions). Toute utilisation commerciale ou impression systématique est constitutive d'une infraction pénale. Toute copie ou impression de ce fichier doit contenir la présente mention de copyright.

NUMDAM
*Article numérisé dans le cadre du programme*
*Numérisation de documents anciens mathématiques*
http://www.numdam.org/

# PRESENTATION D'UNE METHODE DE SEGMENTATION EFFECTUEE A PARTIR DE DEUX FICHIERS DIFFERENTS

## Application à la Sécurité Routière

M. A. CAMBOIS et H. FONTAINE (1)

## I. INTRODUCTION

En matière de sécurité routière, le phénomène accident est si peu fréquent au regard de la population des conducteurs et des kilomètres qu'ils parcourent, qu'il est assez difficile à appréhender directement à partir d'échantillon représentatif des usagers de la route.

On dispose le plus souvent d'une part, d'un fichier recensant les accidents et d'autre part de fichers reflétant la circulation sur le réseau routier ; c'est donc à partir de types de données d'origine tout-à-fait différente que nous effectuons nos études de risque ; nous définissons le risque comme le nombre d'accidents rapportés aux kilomètres parcourus.

L'objet de cet article est d'exposer une méthode permettant d'ordonner les variables disponibles en fonction de leur incidence sur le risque. Ce genre de problème n'est pas spécifique à la sécurité routière et se retrouve très probablement dans d'autres domaines où le phénomène à étudier est rare par rapport à la population susceptible d'être atteinte (dans le domaine de la médecine par exemple).

Nous avons effectué une étude visant à rechercher différents facteurs accidentogènes en passant par la détection d'ensembles de "véhicules x conducteurs x environnement" présentant des risques anormalement élevés en matière de sécurité routière : ceci afin de spécifier la nature des mesures ou actions à entreprendre. Il était souhaitable de faire intervenir sans a priori le plus grand nombre de variables possibles. Pour arriver à ce but, on a procédé de la façon suivante :

On a retenu toutes les variables communes aux deux fichiers reflétant l'un la circulation, l'autre les accidents. Les variables quantitatives ont été rendues qualitatives. On a comparé les distributions de ces variables sur chacun des deux fichiers afin de détecter celles qui font apparaître des risques particulièrement forts ou faibles pour des groupes d'individus donnés. En effet, pour une variable donnée les risques sont d'autant plus hétérogènes que les distributions des accidentés et des circulants sont différentes. Nous allons maintenant détailler le principe de cette méthode.

---

(1) Chargées d'études à l'Organisme National de Sécurité Routière.



*Mots-clés :* **Analyse de données. Segmentation. Hiérarchisation.**

## II. METHODE

Nous avons utilisé une méthode d'analyse statistique dont le principe suit celui de la segmentation. Dans une segmentation classique on cherche à expliquer une variable privilégiée par plusieurs autres variables. Dans notre cas la variable privilégiée n'apparaît pas de manière explicite, elle résulte simplement de la comparaison des distributions des accidentés et des circulants. A part cela, le principe est le même à savoir que l'on cherche à dichotomiser l'ensemble des individus en deux sous-ensembles les plus différenciés possibles et à réitérer cette démarche autant de fois qu'on le juge utile pour l'étude.

### a) Décomposition de chacune des étapes de la segmentation

#### Première étape

L'objectif est de chercher une partition des individus en deux sous-populations les plus hétérogènes possibles entre elles quant à leur risque. Cette partition sera déterminée par la dichotomie de l'ensemble des modalités de la variable qui donne "la distance" la plus élevée entre la distribution des accidentés et la distribution des circulants. La distance entre les distributions peut être définie de différentes manières. Ce point sera détaillé ultérieurement.

Soit p le nombre de variables communes aux 2 fichiers. On appelle $X_i$ l'une de ces variables : i allant de 1 à p, $n_i$ le nombre des modalités de cette variable. Par exemple, la variable Saison aura 4 modalités (Printemps, Eté, Automne, Hiver).

La distribution de l'ensemble des accidentés suivant les $n_i$ modalités de la variable $X_i$ sera :

$$a_{i1}, a_{i2}, \text{-------}, a_{ini}$$

avec

$$\sum_{j=1}^{n_i} a_{ij} = 1$$

Celle de l'ensemble des circulants:

$$c_{i1}, c_{i2}, \text{-------}, c_{ini}$$

avec

$$\sum_{j=1}^{n_j} c_{ij} = 1$$

Etant donné une dichotomie définie par un ensemble K d'indices de la variable $X_i$ et son complémentaire L, deux nouvelles distributions sont obtenues :

– $(a_{iK}, a_{iL})$ chez les accidentés avec

$$a_{iK} = \sum_{j \in K} a_{ij} \text{ et } a_{iL} = \sum_{j \in L} a_{ij}$$

$$(a_{iK} + a_{iL} = 1)$$

– $(c_{iK}, c_{iL})$ chez les circulants avec

$$c_{iK} = \sum_{j \in K} c_{ij} \text{ et } c_{iL} = \sum_{j \in L} c_{ij}$$

$$(c_{iK} + c_{iL} = 1)$$

On calculera une distance entre ces deux nouvelles distributions.

Notons que pour une variable à $n_i$ modalités, il existe $2^{n_i - 1} - 1$ dichotomies possibles, si l'on ne part d'aucun a priori. Ce nombre croît très rapidement avec $n_i$ ; il atteint par exemple pour $n_i = 24$ (le nombre de marques et types) 8 388 607. Nous verrons plus loin comment contourner cette difficulté.

Une distance étant théoriquement calculée pour toutes les dichotomies possibles de toutes les variables, on retiendra la variable et sa dichotomie donnant la distance la plus grande entre les distributions "accidentés" et "circulants". Ceci constituera notre premier critère de segmentation. Il en résultera deux ensembles d'individus à l'intérieur desquels on examinera les statistiques des autres variables relatives aux accidents et à la circulation.

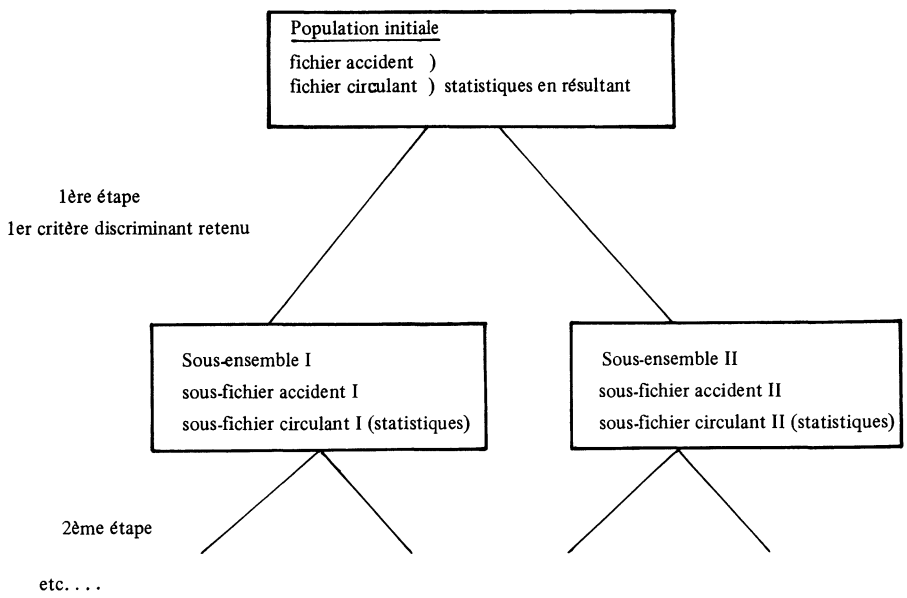
**Deuxième étape**

Sur chacun de ces deux groupes précédemment définis, on répétera le processus de la 1ère étape. Il sera déterminé la nouvelle dichotomie de variable qui soit la plus discriminante pour le risque. On obtiendra alors 4 sous-groupes, sur lesquels de nouvelles statistiques seront calculées.

**$N^{ième}$ étape**

On répétera le même processus pour les $2^{N-1}$ groupes déterminés à l'étape précédente.

On obtient ainsi un arbre donnant une hiérarchie des différents critères pour chacune des branches.



**Règles d'arrêt**

Le traitement étant relativement lourd, deux règles d'arrêt ont été retenues:

– le nombre de pas sera égal au nombre de variables intervenant dans la segmentation, ceci afin de laisser une chance de sortie à chacune d'entre elles (il se peut cependant qu'une même variable apparaisse à des niveaux différents avec une dichotomie différente).

– Dès qu'un groupe déterminé à une certaine étape représentera un pourcentage trop minime d'accidentés ou de circulants, on arrêtera le processus de segmentation sur ce groupe. Il semble a priori qu'une limite inférieure à 10 % soit raisonnable.

**b) Distances proposées**

Soit une dichotomie de la population initiale:

(I, II)

$(a_I, a_{II})$ : la distribution des accidents $(a_I + a_{II} = 1)$

$(c_I, c_{II})$ : la distribution des circulants $(c_I + c_{II} = 1)$

$r_I = \frac{a_I}{c_I}$ : risque relatif du groupe I (rapport des fréquences)

$r_{II} = \frac{a_{II}}{c_{II}}$ : risque relatif du groupe II.

Plusieurs types de distances peuvent être proposées ; nous en utiliserons deux ici ; chacune d'elles conduisant à une interprétation différente de l'arbre en résultant. Il va de soi que l'une d'elles étant choisie, elle sera utilisée exclusivement à chaque pas de la segmentation ; ce qui conduit à deux segmentations possibles.

– *Distance du $\chi^2$ avec la population des circulants en référence*

$$- d_1 = \frac{(a_I - c_I)^2}{c_I} + \frac{(a_{II} - c_{II})^2}{c_{II}} \qquad (1)$$

Lorsqu'il n'y a que 2 classes comme ici, cette formule se simplifie ($a_I = 1 - a_{II}$ et $c_I = 1 - c_{II}$).

On obtient alors :

$$d_1 = \frac{(a_I - c_I)^2}{c_I (1 - c_I)} \qquad (2)$$

En terme de risque, elle peut également prendre la forme suivante :

$$d_1 = c_I (r_I - 1)^2 + c_{II} (r_{II} - 1)^2 \qquad (3)$$

ou encore :

$$d_1 = (1 - r_I)^2 \frac{c_I}{1 - c_I}$$

– *Distance du $\chi^2$ avec la population des accidentés en référence*

$$d_2 = \frac{(a_I - c_I)^2}{a_I} + \frac{(a_{II} - c_{II})^2}{a_{II}}$$

ou encore :

$$d_2 = \frac{(a_I - c_I)^2}{a_I\,(1 - a_I)} = \left(1 - \frac{1}{r_I}\right) \frac{a_I}{1 - a_I}$$

Lors de l'analyse des résultats on examinera la spécificité de chacune de ces distances et les phénomènes qu'elles font apparaître.

**c) Réduction du nombre de distances à calculer à chaque étape**

Nous avons vu, qu'étant donné une variable qualitative définie par un certain nombre de classes, le nombre de dichotomies que l'on peut former croît très vite avec le nombre de classes.

A partir de la variable "saison" par exemple à 4 classes Printemps (P), Eté (E), Automne (A), Hiver (H), on peut former 7 dichotomies possibles :

$$[\{P\} \leftrightarrow \{E,A,H\}\ ;\ [\{H\} \leftrightarrow \{P,A,E\}]\ ;\ [\{E\} \leftrightarrow \{P,A,H\}; [\{A\} \leftrightarrow \{P,E,H\}]$$

$$[\{P,E\} \leftrightarrow \{A,H\}];\ [\{P,A\} \leftrightarrow \{E,H\}]\ ;\ [\{P,H\} \leftrightarrow \{A,E\}]$$

Si nous avons P variables dans la segmentation la ième variable ayant $n_i$ classes, il faudrait former $\sum_{i=1}^{p} (2^{n_i - 1} - 1)$ dichotomies et calculer pour chacune d'elles une distance au 1er pas. Aux pas suivants, le nombre de distance à calculer ne diminue que très légèrement. Ceci rendrait l'exploitation informatique extrêmement lourde et coûteuse.

Nous avons fait le choix de ne pas former toutes les dichotomies possibles. En effet, certaines d'entre elles nous semblent d'emblée peu intéressantes pour ce que l'on recherche.

Prenons l'exemple d'une variable à trois modalités :

– La distribution des accidentés est $\{a_1, a_2, a_3\}$ avec $a_1 + a_2 + a_3 = 1$
– La distribution des circulants est $\{c_1, c_2, c_3\}$ avec $c_1 + c_2 + c_3 = 1$
– Les risques associés à chaque modalité sont $r_1, r_2, r_3$ avec $r_i = a_i/c_i$ pour i allant de 1 à 3.

Classons ces modalités par ordre croissant de risque. Nous obtenons alors $r_I < r_{II} < r_{III}$

La définition du risque est telle que si les deux distributions sont différentes, il y aura toujours un risque supérieur à 1 et toujours un risque inférieur à 1.

$$r_I < 1 \quad a_I < c_I$$
$$r_{III} > 1 \quad a_{III} > c_{III}$$

La dichotomie qui rassemblerait les classes I et III ne ferait que neutraliser ces différences, ce qui est contraire au but recherché. Nous allons donc rejeter

d'emblée ce type de dichotomies pour ne retenir que celles qui préservent l'ordre des risques. En l'occurrence, pour le cas de 3 modalités, il existe toujours une dichotomie préservant l'ordre de risques, soit {I} contre {II, III} soit {I, II} contre {III} qui donnera une valeur de la distance supérieure à celle de la dichotomie {I, III} contre {II}.

Ceci n'a pas été démontré pour une variable à plus de trois modalités, mais nous en avons retenu le principe ; il est d'ailleurs fort probable que dans la majorité des cas, la dichotomie donnant la distance la plus grande se trouve parmi les dichotomies ordonnées.

Nous avons donc procédé de la manière suivante :

Soit une variable qualitative $X_i$ à $n_i$ modalités :

$[a_1, a_2, \ldots, a_{n_i}]$ la distribution des accidents $\left(\sum_1^{n_i} a_j = 1\right)$

$[c_1, c_2, \ldots, c_{n_i}]$ la distribution des circulants $\left(\sum_1^{n_i} c_j = 1\right)$

$[r_1, r_2, \ldots, r_{n_i}]$ le vecteur de risque relatif

avec pour tout j de 1 à $n_i$ $r_j = \dfrac{a_j}{c_j}$

Les classes de la variable $X_i$ vont être ordonnées par ordre croissant de risque :

Soit $[r_I, r_{II}, \ldots . r_{N_i}]$ le vecteur ordonné des risques.

Nous retiendrons la dichotomie donnant la distance la plus grande parmi les $(n_i - 1)$ dichotomies suivantes, formées en coupant en deux la partition ordonnée, la coupure variant de la première à la dernière classe.

$$[\{I\} \leftrightarrow \{II, III, \ldots . N_i\}]$$

$$[\{I,II\} \leftrightarrow \{III,IV, \ldots N_i\}]$$

$$[\{I,II, \ldots, (N_i - 1)\} \leftrightarrow \{N_i\}]$$

Il suffit de calculer les $(n_i - 1)$ distances correspondantes et de retenir la plus élevée.

Ce processus fait passer le nombre de calculs pour chaque variable i de $(2^{n_i - 1} - 1)$ à $(n_i - 1)$.

Le gain est considérable dès que le nombre de modalités est élevé.

## III. APPLICATION PRATIQUE SUR UN EXEMPLE

### III.1. Description des fichiers

Nous avons utilisé les fichiers suivants :

– **Fichier des accidents**

Le fichier d'accidents dont on dispose est le fichier national des accidents corporels qui a été constitué chaque année par le S.E.T.R.A. (1) jusqu'en 1976.

On y trouve des variables concernant l'infrastructure, les intempéries, le véhicule et le conducteur accidentés. Ce fichier est normalement un fichier exhaustif de tous les accidents corporels. Sa taille est très importante, environ 250000 accidents y sont recensés chaque année.

– **Fichier de circulation**

Nous disposons d'un fichier résultat d'une enquête aux stations service qui est un reflet de la circulation de Juillet 1973 à Juin 1974 sur l'ensemble du réseau routier, en agglomération et en rase campagne. La région parisienne en est exclue. Les observations ont été faites entre 8 h et 20 h. Seuls les véhicules légers ont été pris en compte : ils sont au nombre de 8 623. Les données recueillies concernent le véhicule et le conducteur et sont représentatives des kilomètres parcourus.

Le nombre d'accidents correspondant à ces critères étant très élevé, on n'en a retenu qu'un dixième de façon à alléger l'exploitation.

Les variables communes aux deux fichiers sont les suivantes :

**Pour le conducteur** :

– – ***Catégorie socio-professionnelle*** (12 modalités)

. agriculteur, artisan, petit commerçant
. industriel, gros commerçant
. profession libérale, cadre supérieur
. cadre moyen, technicien
. employé de bureau, chauffeur professionnel
. agent de maîtrise, ouvrier
. employé de maison
. membre de l'armée et de la police
. membre du clergé
. étudiant, retraité
. sans profession
. sans réponse

– ***Age du conducteur*** (7 modalités)

. 18-19 ans
. 20-24 ans
. 25-29 ans
. 30-39 ans
. 40-49 ans
. 50-64 ans
. 65 ans et plus

– ***Sexe du conducteur*** (2 modalités)

. Masculin
. Féminin

---

(1) Service d'Etudes Techniques des Routes et Autoroutes.

– *Ancienneté du permis de conduire* (5 modalités)

. 1 à 2 ans
. 3 à 4 ans
. 5 à 9 ans
. 10 à 19 ans
. 20 ans et plus

**Pour le véhicule :**

– *Marque et type* (21 modalités)

Les résultats obtenus à partir de cette variable ayant un caractère confidentiel nous ne le ferons pas apparaître de manière explicite.

– *Année de mise en circulation du véhicule* (5 modalités)

. 1971-1972
. 1969-1970
. 1966-1968
. 1965 et années précédentes

**Date et saison :**

– *Saison* (3 modalités)

. Juin, Juillet, Août, Septembre
. Octobre, Novembre, Décembre, Janvier
. Février, Mars, Avril, Mai

– *Jours* (2 modalités)

. Jours ouvrables
. Week-ends

Cette segmentation nous donnera une hiérarchie des facteurs accidentogènes pour l'ensemble des paramètres concernant le véhicule, le conducteur et la période.

On testera sur ces variables les deux distances proposées afin de mettre en évidence leur caractère propre.

## III.2. Analyse des résultats

**Segmentation utilisant la distance du $\chi^2$ avec les circulants en référence.**

On a obtenu un arbre qui ordonne les variables de l'étude et définit un certain nombre de groupes. Pour chacun de ces groupes l'examen des nouvelles distributions de toutes les variables nous a permis d'approfondir l'analyse de façon à ne pas laisser passer des phénomènes sous-jacents qui pourraient être liés aux paramètres discriminants ; ceci se limitant évidemment aux variables disponibles.

En premier lieu, on peut constater que la distance utilisée à tendance à faire apparaître des groupes tout-à-fait marginaux (1 à 2 % de circulants) ayant par ailleurs des risques relatifs assez élevés. Ainsi, un certain nombre d'étapes de la segmentation a mis en évidence ces groupes ; dans ce cas on a décidé de ne pas accorder trop d'importance à l'ordre d'arrivée de la variable correspondante dans

le classement, sans toutefois perdre de vue que l'émergence de ces groupes peut correspondre à une réalité à ne pas négliger (par exemple le phénomène jeunes conducteurs). La segmentation ne sera pas poursuivie sur des groupes représentant moins de 7 % de la population des circulants.

On peut noter que plus une variable a de modalités, plus le pourcentage de population correspondant à chacune d'elles diminue ; le risque peut ainsi prendre des valeurs assez élevées : en effet, si un groupe représente 1 % des circulants, il suffit que les accidents correspondant représentent 5 % de l'ensemble pour que le risque soit égal à 5 ; alors que pour un groupe formant 20 % des circulants, il faudrait que le pourcentage des accidents corespondant soit égal à 100 pour donner la valeur 5 au risque, ce qui n'arrive jamais.

On trouvera ci-après l'arbre résultant de cette segmentation ainsi que les statistiques concernant la population entière.

On peut en tirer les remarques suivantes :

***a) La première étape fait apparaître la catégorie socio-professionnelle comme variable la plus discriminante pour le risque. Elle distingue deux groupes :***

– d'une part : les agriculteurs, les artisans et petits commerçants, les employés, les chauffeurs professionnels, les agents de maîtrise et les ouvriers, les étudiants, les retraités et les personnes sans profession avec un risque de 1,25. Ce groupe représente 65 % des circulants.

– d'autre part : les industriels, les gros commerçants, les professions libérales, les cadres supérieurs, les cadres moyens et les techniciens, les membres de l'armée, de la police et du clergé avec un risque de 0,54. Ce groupe représente 35 % des circulants.

On peut constater qu'en moyenne le premier groupe a un niveau de revenus inférieur au second. Pour simplifier, nous emploierons la terminologie suivante :

– C S P à bas revenus
– C S P à hauts revenus

Le risque des catégories à bas revenus est plus de deux fois supérieur à celui des catégories à hauts revenus. On va tenter de faire apparaître les phénomènes sous-jacents à cette première partition par l'examen des statistiques dans chacun des deux groupes.

Pour la *population des circulants* on peut noter sur les autres variables les distorsions suivantes :

– Les C S P à hauts revenus ont en moyenne leur permis 3 ans plus tôt que les autres (la moyenne d'âge étant sensiblement la même).

– Les C S P à bas revenus roulent plus fréquemment avec des petites cylindrées et des modèles anciens, l'âge du véhicule étant pratiquement le même.

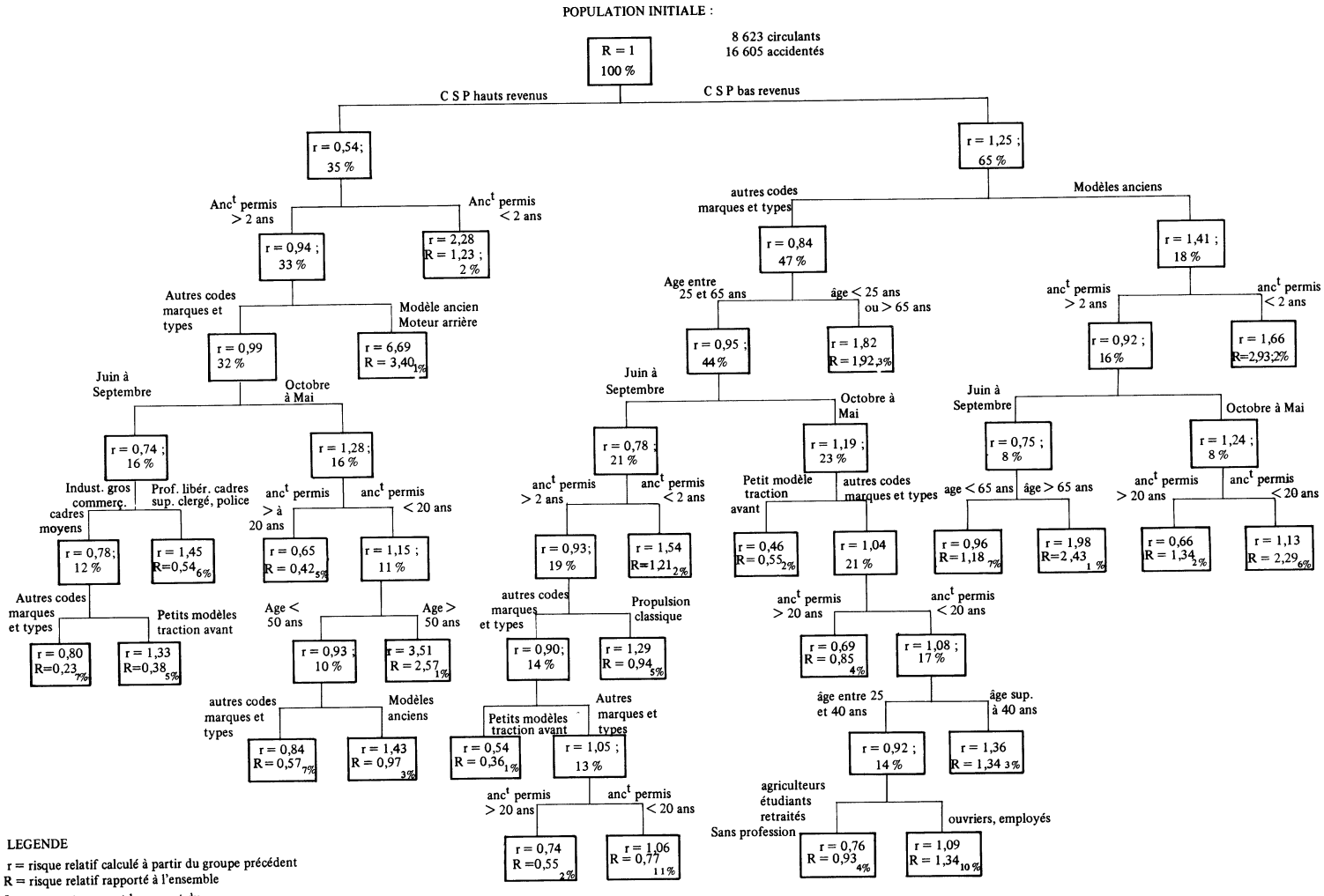
– Les C S P à hauts revenus parcourent proportionnellement plus de kilomètres l'été (Juin à Septembre) que pendant le reste de l'année.

Au vu des statistiques sur l'ensemble de la population, il s'avère que ces remarques sont très liées au risque :

en effet, le risque diminue avec l'ancienneté du permis, les modèles anciens ont un risque plus élevé que les autres et la circulation de Juin à Septembre est relativement moins dangereuse que pendant le reste de l'année.

*Effectifs bruts* : 8 623 circulants 16 605 accidentés

| Variable | Modalités | Circulants | | Accidentés | | Risque |
|---|---|---|---|---|---|---|
| | | Effectif | Pourcentage | Effectif | Pourcentage | |
| *Sexe* | Masculin | 7 160 | 83 % | 13 374 | 81 % | 0,97 |
| | Féminin | 1 463 | 17 % | 3 221 | 19 % | 1,14 |
| Catégorie socio-professionnelle | – Agriculteurs, artisans, petits commerçants | 842 | 10 % | 1 496 | 9 % | 0,95 |
| | – Industriels, gros commerçants | 142 | 2 % | 72 | 1 % | 0,27 |
| | – Professions libérales, cadres supérieurs | 887 | 10 % | 1 143 | 7 % | 0,69 |
| | – Cadres moyens, techniciens | 1 729 | 20 % | 1 534 | 9 % | 0,47 |
| | – Employés de bureau Chauffeurs professionnels | 1 598 | 18 % | 4 033 | 25 % | 1,34 |
| | – Agents de maîtrise, ouvriers | 1 979 | 23 % | 5 025 | 31 % | 1,35 |
| | – Membres de l'armée et de la police | 248 | 3 % | 285 | 2 % | 0,61 |
| | – Membres du Clergé | 34 | 1 % | 45 | 1 % | 0,71 |
| | – Etudiants retraités | 653 | 8 % | 1 512 | 9 % | 1,23 |
| | – Sans profession | 468 | 5 % | 957 | 6 % | 1,09 |
| Age | – 18 - 19 ans | 175 | 2 % | 696 | 4 % | 2,06 |
| | – 20 - 24 ans | 1 564 | 18 % | 3 476 | 21 % | 1,15 |
| | – 25 - 29 ans | 1 782 | 20 % | 2 742 | 17 % | 0,83 |
| | – 30 - 39 ans | 2 077 | 24 % | 3 572 | 21 % | 0,89 |
| | – 40 - 49 ans | 1 686 | 20 % | 2 948 | 18 % | 0,90 |
| | – 50 - 64 ans | 1 135 | 13 % | 2 313 | 14 % | 1,05 |
| | – plus de 64 ans | 230 | 3 % | 819 | 5 % | 1,84 |
| Ancienneté du permis | – 1 - 2 ans | 936 | 11 % | 3 128 | 19 % | 1,72 |
| | – 3 - 4 ans | 957 | 11 % | 2 063 | 12 % | 1,11 |
| | – 5 - 9 ans | 2 108 | 25 % | 4 015 | 24 % | 0,98 |
| | – 10 - 19 ans | 2 601 | 30 % | 4 956 | 30 % | 0,98 |
| | 20 ans et plus | 1 932 | 23 % | 2 443 | 15 % | 0,65 |
| Ancienneté du véhicule | – 73 - 74 | 1 732 | 20 % | 2 726 | 17 % | 0,82 |
| | – 71 - 72 | 2 245 | 26 % | 3 876 | 23 % | 0,90 |
| | – 69 - 70 | 1 693 | 20 % | 3 035 | 18 % | 0,93 |
| | – 66 - 68 | 1 694 | 20 % | 3 631 | 22 % | 1,11 |
| | – avant 1966 | 1 259 | 14 % | 3 337 | 20 % | 1,38 |
| Quadrimestre | – Juin à Septembre | 4 246 | 49 % | 6 205 | 37 % | 0,76 |
| | – Octobre à Janvier | 1 868 | 22 % | 4 896 | 30 % | 1,36 |
| | – Février à Mai | 2 509 | 29 % | 5 504 | 33 % | 1,14 |
| Type de Jour | – Ouvrable | 5 448 | 63 % | 10 525 | 63 % | 1,00 |
| | – Férié | 3 175 | 37 % | 6 080 | 37 % | 0,99 |




LEGENDE

r = risque relatif calculé à partir du groupe précédent
R = risque relatif rapporté à l'ensemble

Les pourcentages sont le rapport du nombre de circulants du groupe au nombre de circulants dans la population initiale

SEGMENTATION SUR LES VARIABLES CROISEES "VEHICULES x CONDUCTEURS"

– Distance du $\chi^2$ avec les circulants en référence

*b) La deuxième étape fait apparaître deux nouvelles variables :*

– L'ancienneté du permis pour les C S P à hauts revenus en isolant les débutants avec un risque 2,3 fois plus fort que les autres. Bien qu'en soi le manque d'expérience soit un facteur de risque, il se trouve par ailleurs que ce groupe formé d'individus jeunes (en moyenne 23 ans) conduit relativement plus de petites cylindrées et d'anciens modèles réputés comme des véhicules moins sûrs.

– Les marques et types pour les C S P à bas revenus isolant les modèles avec un risque 1,7 fois plus élevé que les autres. On ne distingue pas d'autres différences notables entre les deux groupes ainsi formés.

*c*) La troisième étape est complémentaire de la seconde, dans la mesure où l'on retrouve les mêmes variables inversées : ainsi, dans les C S P à hauts revenus la marque et type succède à l'ancienneté du permis

Dans les C S P à bas revenus l'âge et l'ancienneté du permis qui sont des variables très liées, succèdent à la marque et au type. Dans le groupe des véhicules à haut risque les débutants sont isolés comme sous-groupe plus dangereux ; pour les autres véhicules, on distingue les conducteurs de moins de 25 ans et de plus de 65 ans présentant un risque plus élevé que les autres.

*d*) Quatrième étape. Il est remarquable qu'à ce niveau, quel que soit le groupe dans lequel on se trouve, c'est la période de l'année qui apparaît, en isolant à chaque fois les 4 mois d'été (Juin à Septembre) comme période plus sûre que le reste de l'année, les risques relatifs étant tout-à-fait comparables, les mois les plus froids sont 1,5 à 1,7 fois plus dangereux que les mois d'été.

*e)* Aux étapes suivantes on retrouve les mêmes variables : âge, ancienneté du permis, marque et type et catégorie socio professionnelle avec d'autres modalités segmentantes. Les conducteurs ayant leur permis depuis plus de 20 ans apparaissent plusieurs fois comme groupe particulièrement sûr. Inversement, les conducteurs âgés de plus de 50 ans ou 65 ans apparaissent comme groupe plus dangereux.

A aucun moment, les variables suivantes ne sont apparues directement comme discriminantes :

– l'âge du véhicule
– sexe
– jour ouvrable, week-end

Ceci n'est pas étonnant dans le cas des deux dernières variables où le risque est très proche de 1, quelles que soient les modalités (hommes : 0,97 et femmes : 1,14 ; jours ouvrables : 1,00 et week-end : 0,99).

Dans le cas de la première variable, le risque croît régulièrement avec l'ancienneté du véhicule, tout en restant dans une fourchette assez restreinte (moins d'un an : 0,81 ; plus de 8 ans : 1,38).

Nous mentionnons que nous avons également effectué une autre segmentation en utilisant la distance du $\chi^2$ avec les accidentés en référence. Nous avons noté que l'ordre d'arrivée des variables était le même et que la tendance de cette distance était d'isoler les groupes marginaux à risque particulièrement faible. Ceci n'est pas étonnant étant donné la manière dont peut s'exprimer cette distance en fonction du risque (cf. § II.b).

*Remarque* : Nous disposions également d'une autre enquête reflétant la circulation. Les observations étaient faites au bord de la route sans arrêter les véhicules.

Les données recueillies concernent l'infrastructure, les conditions climatiques et les types de véhicules. Une segmentation a été effectuée à partir de ce fichier de circulation et du fichier accident. On a été confronté ici à des problèmes d'homogénéité de codification d'un fichier à l'autre plus importants que précédemment. En effet, le recueil de données implique une normalisation des codifications, en particulier pour les variables de météorologie et d'infrastructure (force du vent, déclivité de la chaussée). Il est donc apparu des distorsions dans les distributions de ces variables tout à fait artificielles et ne correspondant pas à des risques réels. Il est primordial avant d'effectuer ce type d'analyse de s'assurer de l'homogénéité de la codification des les deux fichiers utilisés.

## IV. CONCLUSION

Comme toute segmentation, ce type d'analyse est assez lourd et coûteux malgré le choix que nous avons fait de réduire le nombre de dichotomies à prendre en compte. Un autre inconvénient s'est ajouté ici dans la mesure où il a fallu veiller à ce que les systèmes de codification soient compatibles d'un fichier à l'autre. Mais une segmentation classique effectuée à partir d'un fichier unique réunissant pour chaque individu des données sur ses caractéristiques habituelles et sur ses accidents s'il y a lieu, se heurte à d'autres problèmes tout aussi délicats, liés à la rareté du phénomène étudié. Dans ce cas, la méthode que nous avons mise au point peut constituer un outil statistique plus adapté.