MINISTÈRE DE LA JUSTICE
DIRECTION DE L'ADMINISTRATION PÉNITENTIAIRE
CENTRE NATIONAL D'ÉTUDES ET DE RECHERCHES PÉNITENTIAIRES

Recherche sur une prévision des effectifs de la population pénale

(éléments de méthode et premiers résultats)

1971

Recherche sur une prévision des effectifs de la population pénale

(éléments de méthode et premiers résultats)

Recherche sur une prévision des effectifs de la population pénale

(éléments de méthode et premiers résultats)

L'administration pénitentiaire est appelée à faire face à de fréquentes et parfois importantes variations dans le volume de la population des prisons, liées au mouvement de la délinquance et à l'activité des services judiciaires et de police. Il importe par suite de réaliser la meilleure approche statistique possible des tendances de l'évolution de la population pénale en vue, notamment, d'adapter les équipements aux besoins à satisfaire. C'est pourquoi le Centre national d'Etudes et de Recherches pénitentiaires a été appelé à entreprendre, en 1969, une recherche sur la prévision des effectifs des détenus dans les années à venir (1). Ce programme était justifié, indépendamment de son intérêt scientifique, par la nécessité d'évaluer l'importance des équipements à prévoir pour la période du VI[e] Plan (1970 - 1975).

Il importe, avant d'exposer la méthodologie et les premiers résultats de ce programme, de faire deux remarques préalables.

Les études prévisionnelles appliquées à la criminalité sont récentes, et il n'existe pas en particulier de précédent d'une recherche de ce type concernant la population des prisons : le choix d'une méthodologie s'est donc heurté à un certain nombre de difficultés.

La nécessité, en second lieu, de fonder la prévision non pas sur une simple prolongation de la tendance en cours — qui a d'ailleurs fait l'objet d'une prérecherche (2) — mais sur une analyse plus sophistiquée de la criminalité a conduit à instituer une coopération entre les trois unités de recherche criminologique du ministère de la Justice en vue de l'aboutissement d'un programme coordonné.

---

(1) Equipe composée de A. MORINEAU, G. PICCA, O. RABUT, P. VENGEON, A. BEAUCHESNE. Les recherches et définitions des modèles ont été conduites par A. MORINEAU, docteur ès sciences.

(2) C.N.E.R.P., *Prérecherche sur les perspectives d'évolution de la population des prisons dans les années à venir* (réflexions sur 1971, 1975 et 1985), ronéo, 1968.

L'objet du présent rapport est moins de rendre compte de l'ensemble de ce programme — qui a déjà fait l'objet d'une publication (3) — que d'exposer la méthode suivie concernant la prévision de la population des prisons et les premiers résultats obtenus dans ce domaine. Il importe néanmoins, à cette occasion, de souligner l'utilité de la coopération instituée entre les trois unités de recherche, et notamment la contribution apportée par le Service d'Etudes Pénales et Criminologiques de la direction des affaires criminelles dans la collecte des données statistiques et leur interprétation, et celle de la section informatique du Centre de Recherches de l'Education Surveillée dans le traitement sur ordinateur des informations. Ces contributions se sont, en effet, révélées décisives pour l'aboutissement des travaux dont il va être rendu compte.

* * *

L'analyse des premiers résultats figurant dans ce rapport permet de formuler un certain nombre d'hypothèses sur le volume des effectifs de prévenus et de condamnés prévisible en 1975. Ces chiffres doivent toutefois être appréciés en fonction de deux facteurs qui n'ont pu être pris en compte dans cette recherche.

En effet, il était nécessaire sur un plan technique d'établir une liaison statistique entre les chiffres de la criminalité légale et ceux de la population pénale. Or, les derniers éléments d'information publiés sur la criminalité remontent à 1968 ; les projections pour 1975 ont donc pu être établies à partir de cette année de référence et non de 1970. Dès lors, le calcul prévisionnel n'a pu prendre en compte la nette diminution d'effectif de la population pénale *observée* en 1969.

Par ailleurs, il y a lieu de prendre en considération, non seulement une certaine instabilité dans le volume des différentes catégories de la population pénale, mais aussi, plus généralement, l'influence des modifications de la législation sur l'évolution de ce volume. C'est ainsi qu'on peut légitimement espérer que les dispositions de la loi du 17 juillet 1970 pourront entraîner un certain ralentissement dans le nombre des incarcérations.

Ces deux correctifs sont de nature à réduire dans une proportion de 5 à 6 % les résultats présentés dans ce rapport.

---

(3) PICCA (G.) ET ROBERT (PH.), « Notes sur une recherche prévisionnelle de l'évolution de la criminalité », *Revue française de sociologie*, XI, 1970, p. 390-405.

S O M M A I R E

I. — INTRODUCTION

1. — Le raisonnement sur modèle en criminologie

Toute réflexion sur le phénomène criminel repose sur une série d'observations (qualitatives ou chiffrées) destinées à apporter une certaine connaissance du phénomène. On appelle MODÈLE une REPRÉSENTATION FORMELLE de cette connaissance, c'est-à-dire une écriture sous forme de système mathématique.

Il n'est pas de science qui n'utilise le modèle comme instrument de raisonnement, la démarche consistant à explorer les conséquences logiques des hypothèses du modèle pour les confronter aux observations et par là appréhender la réalité du phénomène. L'ensemble de ces représentations abstraites constitue le point de départ inéluctable de toute recherche scientifique. Mais alors que dans les sciences physiques on peut atteindre des représentations fidèles et souvent exactes, la situation est bien différente dans les sciences humaines et sociales et notamment en criminologie. Car la criminalité est intégrée dans le contexte mouvant de l'organisation sociale. On ne pourra, par suite, espérer atteindre que des représentations approximatives obtenues après de nécessaires simplifications. Néanmoins, l'introduction du raisonnement logique sur un modèle même imparfait est la condition sine qua non de la rigueur en même temps que l'expression d'une certaine modestie du chercheur conscient des limites de sa compréhension. En d'autres termes, la méthode scientifique en criminologie, comme dans toute science sociale, doit passer par cette porte étroite que constitue le MODÈLE.

Un modèle explicite les relations existant entre certaines variables caractérisant le phénomène étudié dans son contexte. Ces grandeurs peuvent être classées en deux groupes, les *variables ENDOGÈNES* qu'on considère comme déterminées par le phénomène étudié (par exemple, pour un modèle criminel, les effectifs criminels dans les diverses catégories d'infractions), et les *variables EXOGÈNES* qui correspondent à toutes les autres variables intervenant dans les relations du modèle (indicateurs économiques, sociaux ou démographiques, etc.). Le modèle explicite comment les variables exogènes déterminent les variables endogènes.

Il est important toutefois de souligner que la dépendance entre variables endogènes et exogènes *n'est pas nécessairement de nature causale.* Lorsqu'en physique on explicite le modèle liant la pression d'un gaz à son volume à température constante et qu'on écrit que la pression est inversement proportionnelle au volume occupé par le gaz, on n'en conclut pas que le volume est la CAUSE directe de la pression

observée sur les appareils enregistreurs (1). Il faudra se souvenir de cette remarque élémentaire lorsqu'on parlera plus loin des modèles linéaires de la criminalité. En effet, les modèles que l'on utilise le plus souvent sont des *MODELES DE SIMULATION*, modèles qui reviennent à fonder le raisonnement sur la constatation : « tout se passe comme si ... ». Un modèle qui traduit des relations causales est en particulier un modèle de simulation, mais la réciproque est fausse en général. Pour la plupart des conséquences logiques que l'on veut déduire des raisonnements sur les modèles, les modèles de simulations suffisent ; de plus c'est souvent la détermination de modèles de simulation qui concourt à l'élaboration ultérieure de modèles de causalité.

Pour résumer ces remarques relatives à la théorie des modèles appliqués à la criminologie, on pourrait très schématiquement dire que le raisonnement sur un modèle est une nécessité, et que les réflexions des criminologues au cours des temps ont porté en général sur l'analyse des observations préliminaires à l'écriture d'un modèle causal formalisé (le modèle lui-même ayant rarement été écrit). Dans le présent rapport il ne sera pas entrepris l'écriture d'un modèle causal formalisant une certaine théorie mais plutôt l'explicitation d'un modèle de simulation.

2. — Induction statistique et modèles « A PRIORI »

Pour évoquer le rôle fondamental du modèle dans la recherche criminologique on a parlé au paragraphe précédent d'une liaison purement fonctionnelle entre les variables exogènes et endogènes. Plus généralement on peut dire que les modèles contiennent des éléments aléatoires, aucune relation fonctionnelle ne pouvant s'ajuster exactement aux observations faites. Dans ce cas, le modèle ne détermine pas les variables endogènes en fonction des variables exogènes, mais définit plutôt la LOI DE PROBABILITÉ des variables endogènes, loi de probabilité conditionnée par les valeurs des variables exogènes. L'introduction de l'aléatoire sera évidemment fondamentale en criminologie où les représentations théoriques risquent de s'écarter des observations bien davantage que dans les sciences physiques. D'autre part, c'est l'aléatoire qui va conditionner évidemment la nature des procédures statistiques, la mesure de leur adéquation, et l'interprétation de leurs résultats.

En effet, la réflexion sur les observations autorise en général à proposer un type de modèle a priori (2) sans toutefois permettre de le spécifier complètement. Compte tenu des faits observés, c'est l'analyse statistique qui va permettre de préciser le modèle et d'en écarter certaines formes jugées peu vraisemblables. Il y a cependant une règle logique fondamentale à rappeler à propos de l'induction statistique :

---

(1) On dispose cependant d'un modèle « causal » où la pression est expliquée par le choc des particules gazeuses sur les parois (nombre et vitesse moyenne des particules).

(2) A PRIORI ne signifie pas ici « choisi n'importe comment ». Tout au contraire on insistera plus loin sur les précautions à prendre dans le choix du modèle « a priori ».

LES MÉTHODES DE LA STATISTIQUE MATHÉMATIQUE NE PERMETTENT JAMAIS DE SPÉCIFIER LA NATURE D'UN MODÈLE. Le seul propos de l'induction statistique est de préciser un modèle *choisi auparavant et a priori*, ou ce qui revient au même de choisir entre plusieurs modèles qui soient des cas particuliers D'UN MÊME modèle a priori choisi auparavant. Dans tous les cas, l'induction statistique s'applique toujours à un modèle a priori qui ne peut jamais être remis en question, soit pour déterminer les paramètres du modèle, soit pour faire choix d'une forme particulière de ce modèle a priori. Ainsi par exemple, aucune méthode statistique ne permet de dire si une liaison entre deux variables est linéaire plutôt que logarithmique, à moins que ces deux natures de liaisons puissent apparaître comme deux cas particuliers d'une fonctionnelle plus générale : il faut alors nécessairement supposer que cette fonctionnelle représente la nature RÉELLE de la liaison pour pouvoir décider si la liaison est linéaire ou logarithmique ; de toute façon *il y a toujours un modèle choisi a priori comme représentant correctement le phénomène réel.*

Remarquons à ce propos que si l'on voulait déterminer « scientifiquement » un modèle causal de la criminalité, il faudrait rassembler les théories littéraires les plus vraisemblables qui ont été émises au cours du temps et établir, si c'est possible, une formulation abstraite d'un modèle général qui admette comme cas particuliers les formulations des diverses théories retenues. Alors, et alors seulement, l'induction statistique permettrait d'effectuer le choix d'une de ces théories comme étant la mieux adaptée aux observations. Signalons enfin que pour pouvoir conclure plus radicalement que la théorie sélectionnée par l'induction statistique est la « vraie » théorie, celle qui représente le mécanisme réel du phénomène criminalité, il faudrait être assuré que la formulation générale utilisée est susceptible de représenter le phénomène réel — certitude qui sera évidemment toujours hors d'atteinte (1). On comprend par suite d'autant plus aisément que notre ambition se limite ici à la recherche de modèles de simulation, qui seront d'ailleurs suffisants pour notre objectif.

3. — Recherche de modèles : analyse des données

On a rappelé plus haut l'exigence absolue d'un modèle choisi a priori comme représentant la liaison RÉELLE entre les variables exogènes et endogènes si l'on veut s'assurer d'une certaine rigueur dans le raisonnement. On conçoit, par suite, que le choix de ce modèle soit déterminant puisque tout reposera finalement sur l'hypothèse INCONTROLABLE et posée au départ comme axiome que le modèle utilisé est correct. Par conséquent, les plus grandes précautions doivent être prises dans le choix de ce modèle. Au stade où se trouve la science criminologique on peut même affirmer que la recherche des modèles a priori devrait être sa préoccupation principale.

---

(1) A titre de consolation signalons que TOUTES les sciences sociales connaissent le même problème.

Les difficultés qui apparaissent dans la recherche des modèles a priori sont liées à la nature des observations dont on dispose pour les étayer.

Il y a d'une part une telle abondance et une telle diversité d'observations et de statistiques concernant l'émergence de la criminalité dans la société, qu'il est souvent impossible à l'esprit de saisir immédiatement ce que traduisent les observations. Par ailleurs, ces observations sont généralement le résultat de collectes routinières plutôt que le fruit d'expérimentations contrôlées, de telle sorte que le chercheur ne peut pas être maître des conditions d'invariance des observations. Ces difficultés ne sont pas spécifiques à la criminologie et sont inhérentes à la plupart des sciences humaines (économétrie, sociométrie, etc.). Elles rendent l'induction statistique particulièrement délicate. Aussi pour tirer des nombreuses observations disponibles une certaine connaissance du phénomène propre à conduire au choix d'un modèle a priori sur lequel s'exercera l'induction statistique, il est particulièrement précieux de pouvoir synthétiser l'ensemble des informations. Une branche de la statistique moderne qui doit son développement récent aux progrès du calcul automatique, l'ANALYSE DES DONNÉES, s'avère particulièrement efficace dans cette entreprise :

Le principe en est relativement simple. Les informations sont rassemblées dans un tableau numérique et sous cette forme s'interprètent comme un « nuage de points » dans un hyper-espace. Ce nuage en général n'est pas « sphérique » mais s'allonge plus particulièrement dans certaines directions privilégiées correspondant évidemment à des propriétés de dépendance ou d'association des observations. Ce sont ces liaisons, en général cachées par la surabondance des informations, que l'analyse des données permet d'extraire, en même temps qu'elle en permet généralement l'identification grâce à des représentations graphiques synthétiques faciles à lire. Suivant la nature du tableau des données statistiques on fait choix d'une certaine définition des distances entre les points du nuage pour déterminer les directions d'allongement, ce qui donne lieu à diverses formes d'analyse des données : analyse en COMPOSANTES PRINCIPALES, analyse des CORRESPONDANCES, analyse des RANGS, analyse des covariances partielles, etc. Dans tous les cas, et contrairement aux méthodes de l'induction statistique, l'analyse des données est indépendante de tout modèle ou de toute hypothèse a priori sur le phénomène. Elle permet de décrire et de synthétiser de façon strictement objective un ensemble complexe d'informations. C'est pourquoi on pourra l'utiliser sans réserve pour extraire des observations les idées ou les connaissances qui permettront d'écrire les modèles a priori sur lesquels s'exercera l'induction statistique.

4. — Problème général des prévisions

La prévision sera considérée ici comme une opération logique puisant ses méthodes dans l'induction statistique. Autrement dit il ne sera pas fait état de la prévision « intuitive » trouvant son fondement dans les jugements d'experts, et institutionnalisée par exemple dans la technique dite des « scénarios » mise en œuvre dans certaines administrations. L'opération de prévision s'effectue classiquement en deux étapes ; la première est l'application de l'induction statistique pour préciser complètement un modèle a priori de simulation (ou de causalité) du phénomène étudié dont on vient de donner ci-dessus les principes fondamentaux. La seconde est une nouvelle application de l'induction statistique pour utiliser correctement le modèle à des fins prédictives. Il s'agit là d'une branche technique de la statistique mathématique. On signalera, au moment opportun, les propriétés classiques que nous devrons utiliser à propos des deux modèles développés : le « modèle linéaire » de la criminalité, et le « modèle à retards échelonnés » du passage de la population criminelle à la population pénale.

Quoi qu'il en soit des problèmes et des solutions techniques, il faut se souvenir qu'une prévision est toujours et nécessairement CONDITIONNELLE, ceci à deux titres indépendants. Tout d'abord la prévision est conditionnée par l'adéquation du modèle qu'il a fallu nécessairement choisir a priori comme support du raisonnement ; on ne pourra pas lui accorder plus de confiance qu'on en accordera au modèle lui-même (puisqu'il est impossible d'avoir la certitude que le modèle est correct (1). D'autre part, la prévision est conditionnée par les valeurs prises par les variables exogènes puisque l'induction statistique détermine la loi de probabilité conditionnelle des variables endogènes. Ici encore la confiance qu'on peut accorder à une prévision est limitée par les erreurs de mesure sur les variables exogènes, en particulier lorsque celles-ci sont elles-mêmes des prévisions.

Signalons au passage qu'on vient d'identifier deux sources d'erreurs sur les prévisions. Dans certains cas l'erreur due aux mesures sur les variables exogènes peut être estimée si l'on connaît la loi des erreurs de mesure (par exemple des distributions de probabilité sur les variables exogènes quand il s'agit de prévisions) ; mais l'erreur due à l'utilisation d'un modèle incorrect n'est évidemment jamais identifiable ni estimable puisqu'on ne connaît pas le « vrai » modèle. Ces erreurs sont dues aux ERREURS DE SPÉCIFICATION du modèle et la théorie montre que les erreurs de spécification peuvent avoir un impact important sur les prévisions. Une méthode empirique, mais souvent efficace, pour éliminer des erreurs de spécification grossières consiste à construire plusieurs modèles a priori assez peu différents les uns des autres, comme autant d'approximations du modèle véritable inconnu. On se trouve certainement proche du modèle correct si les prévisions sont stables dans les différents modèles. C'est cette méthode qu'on préconisera dans l'étude du modèle linéaire de la criminalité. Il reste évidemment, outre les erreurs possibles qu'on vient de citer, un autre type d'imprécision concernant les prévisions, à savoir, leur nature aléatoire dans un modèle aléatoire. En effet, ce qu'on connaît finalement ou du moins ce qu'on peut en général estimer par induction statistique c'est la loi de probabilité conditionnelle de la

(1) On connaît actuellement quelques méthodes de « prévisions directes » c'est-à-dire ne nécessitant pas la spécification d'un modèle a priori. Mais ces méthodes sont encore très frustes, limitées et peu souples.

quantité à prévoir. Si on veut retenir une valeur précise, il convient alors de lui attribuer un certain coefficient d'incertitude (par exemple un intervalle de confiance correspondant à un certain seuil).

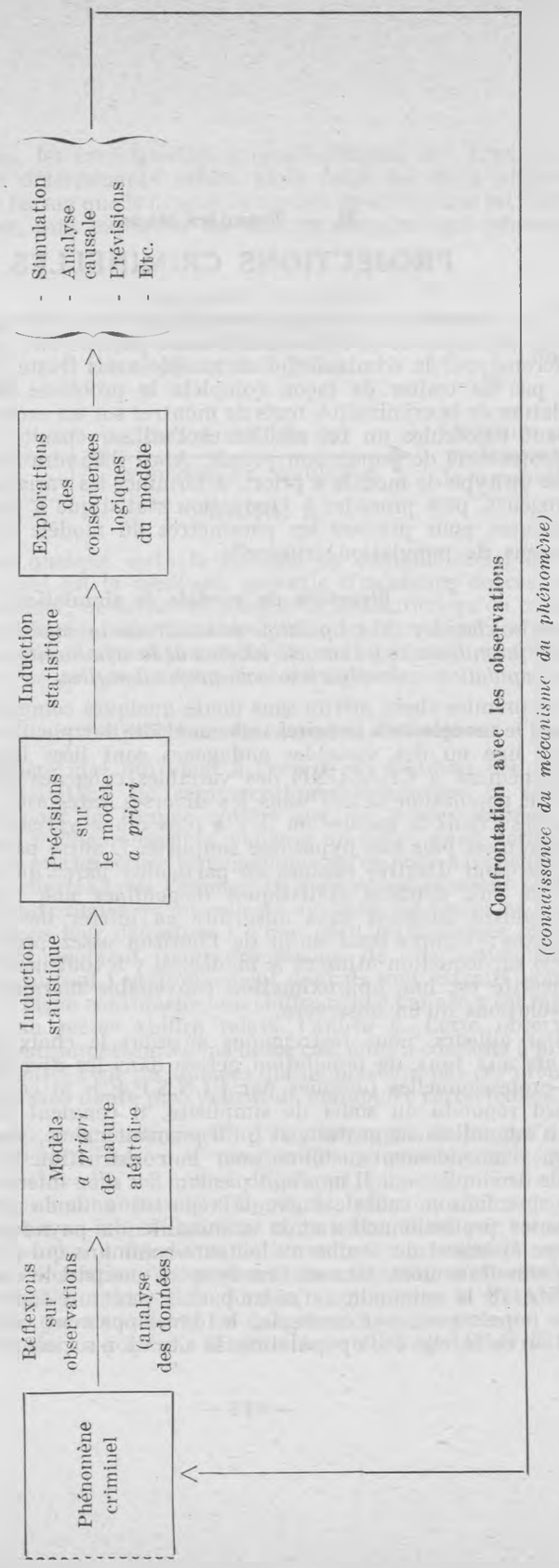
5. — Brèves conclusions

Le problème à résoudre était celui de la prévision à moyen terme du volume de la population pénale ; les grandes étapes successives de la méthode que l'on propose vont apparaître comme autant d'applications des remarques générales qui viennent d'être rappelées.

Tout d'abord on cherchera un modèle a priori formalisant le processus complexe (procédures judiciaires) par lequel la criminalité alimente les prisons. Il s'agira d'un modèle de simulation, et non d'un modèle causal ou descriptif suivant toutes les étapes du processus. La nature de ce modèle a priori sera fixée par des considérations logiques et l'observation de séries chronologiques d'effectifs (1). L'induction statistique permettra de PRÉCISER les paramètres de ce modèle, où la variable endogène est la population pénale, et la variable exogène la population criminelle. C'est aussi l'induction statistique qui permettra d'utiliser ce modèle pour établir une PRÉVISION de population pénale conditionnée par la population criminelle attendue. On comprendra par là même qu'il était indispensable d'établir une prévision de population criminelle comme étape intermédiaire de la recherche. Afin de débrouiller l'écheveau de données très abondantes et complexes, on aura recours aussi souvent que nécessaire à des ANALYSES DE DONNÉES. Ces analyses permettront de choisir les variables exogènes d'un MODÈLE DE SIMULATION supposé a priori le plus simple, c'est-à-dire linéaire, et faisant intervenir exclusivement les caractéristiques socio-économiques et démographiques de la société. L'induction statistique conduira ensuite à l'estimation des paramètres du modèle et à son utilisation comme instrument de prévision de la criminalité. Le cheminement inéluctable du raisonnement scientifique est synthétisé sur le schéma n° 1.

(1) Il s'agit d'un modèle dit à retards échelonnés ou « distributed lag » dans la littérature anglo-saxonne.

FIGURE 1



II. — Première étape

PROJECTIONS CRIMINELLES

Il est très important de garder en mémoire pour ce chapitre d'une part les principes de la recherche, d'autre part, le fait que nous nous satisferons pour la criminalité d'un modèle assez fruste ; notre propos n'est pas de traiter de façon complète le problème du modèle de simulation de la criminalité, mais de montrer sur un exemple comment on peut construire un tel modèle et l'utiliser ensuite pour obtenir des projections de population pénale. Ainsi il faudra successivement définir un type de modèle a priori, déterminer les variables endogènes et exogènes, puis procéder à l'induction statistique à partir de séries de données pour préciser les paramètres du modèle et calculer les prévisions de population criminelle.

1. — Structure du modèle de simulation

On va chercher 'il est possible de construire un modèle de simulation liant de façon linéaire les caractéristiques de la criminalité à la répartition de la population en catégories socio-professionnelles.

Ce premier choix mérite sans doute quelques commentaires. Tout d'abord le modèle sera *linéaire*, autrement dit il explicitera dans quelle mesure une ou des variables endogènes sont liées linéairement et simultanément à CHACUNE des variables exogènes que seront les taux de population active dans les diverses catégories socio-professionnelles ; dans la mesure où il y a plus d'une variable exogène, la linéarité n'est plus une hypothèse simpliste. D'autre part, la linéarité s'impose pour d'autres raisons, en particulier parce qu'on ne saurait accorder aux données statistiques disponibles une confiance telle qu'on puisse imaginer sans absurdité en inférer des modèles plus complexes ; compte tenu enfin de l'horizon assez proche (quelques années) sur lequel on utilisera le modèle, il y a tout lieu de croire que la linéarité est une approximation convenable marginalement pour les évolutions qu'on observera.

Par ailleurs, nous restreignons a priori le choix des *variables exogènes* aux taux de population active dans les diverses catégories socio-professionnelles (définies par l'I.N.S.E.E.). Si ce choix a tout d'abord répondu au souci de simplicité, il convient de remarquer qu'il n'est nullement gratuit, et qu'il pourrait même, sous réserve de quelques amendements, suffire pour la construction d'un excellent modèle de simulation. Il ne s'agit pas en effet d'en inférer qu'il peut y avoir une liaison causale entre la répartition de la population en catégories professionnelles et la criminalité du pays ; cependant on imagine aisément de nombreux facteurs communs qui peuvent déterminer simultanément dans un lieu donné d'une part le « volume » et la « qualité » de la criminalité d'autre part la structure socio-économique de sa population ; par exemple, le développement économique, la situation culturelle de la population, la situation sociale prépondérante

des individus, les caractéristiques géographiques, etc. L'existence de ces facteurs déterminants assure alors entre les deux phénomènes une certaine liaison que le propos du modèle de simulation est justement de quantifier, sans expliciter les chaînes causales (qui peuvent être complexes).

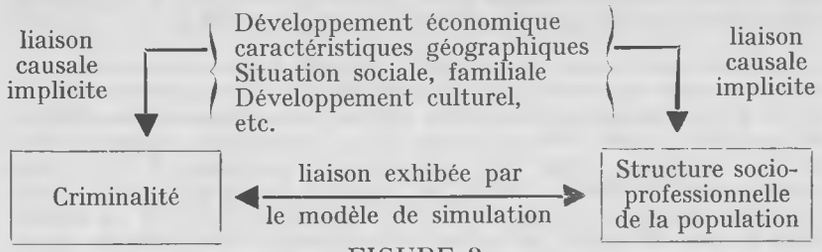


FIGURE 2

C'est en quelque sorte la richesse de contenu social des deux phénomènes qui est la meilleure garantie d'existence de ces liaisons implicites, et donc, de succès du modèle de simulation qu'on recherche. Bien que ce commentaire soit trop bref, on se convaincra aisément à partir de là que ces premiers choix ne sont pas aussi simplistes ou restrictifs qu'ils peuvent apparaître dès l'abord.

2. — Le problème des données statistiques

La première idée qui vient à l'esprit est d'utiliser des SÉRIES CHRONOLOGIQUES de caractéristiques criminelles et socio-professionnelles pour la France entière sur les années passées. Cette procédure est malheureusement vouée à un échec quasi certain, non pas tant parce que les séries chronologiques qu'on pourra recueillir seront relativement courtes (voir l'annexe I), mais essentiellement parce que les données statistiques qu'on collecte au cours du temps ne sont pas homogènes dans leur définition : d'une part, les principes mêmes de dénombrement évoluent (pour des raisons de « perfectionnement »), d'autre part, surtout l'environnement dont ils sont extraits se modifie également de façon continuelle ; un chiffre relevé l'année x est rarement comparable au même chiffre relevé l'année y. Cette observation, particulièrement importante dans notre cas, nous a conduits à proposer d'utiliser un autre type de données qui ne présente pas ces problèmes, mais qu'il est sans doute plus délicat de manipuler correctement.

La France a la particularité d'être découpée de longue date en DÉPARTEMENTS tels que chacun d'eux possède une certaine « personnalité » — plus ou moins développé, plus ou moins commerçant, plus ou moins peuplé d'immigrés, plus ou moins jeune de population, plus ou moins chaud de climat, etc. Il est par conséquent tentant de chercher à déceler la liaison entre criminalité et variables exogènes de simulation en prenant comme unité d'observation le DÉPARTEMENT à une date donnée, et non plus la FRANCE au cours des ans qui s'écoulent. Le danger sera évidemment d'affirmer ex-abrupto, pour une liaison déterminée, qu'il s'agit là de la liaison

d'évolution temporelle ; pour pouvoir énoncer une telle affirmation, il est clair qu'il sera nécessaire de prendre de nombreuses précautions. D'où le principe conducteur :

L'induction statistique sera effectuée sur des données DÉPARTEMENTALES (à une date donnée) pourvu qu'il soit possible d'exhiber une liaison « départementale » qui coïncide avec ce que l'on connaît de l'évolution temporelle entre les variables exogènes et endogènes du modèle.

Entendons-nous, le problème *n'est pas* de ranger les départements dans un ordre tel qu'on puisse dire que chacun d'eux représente l'état de la France entière à des dates successives : Haute-Loire 1968 = France 1962 ; Loiret 1968 = France 1963 ; Yonne 1968 = France 1964, etc. Ce problème serait très certainement insoluble. Il s'agit très précisément d'exhiber parmi les liaisons départementales entre variables exogènes et endogènes celles qui coïncident, s'il en existe, avec ce qui est connu de la liaison temporelle. Une première façon d'éliminer des liaisons départementales qui ne conviennent pas est évidemment d'utiliser le modèle obtenu en « simulation », c'est-à-dire de lui faire prédire une certaine valeur déjà observée des variables endogènes : si l'écart entre valeurs observées et valeurs calculées est grand, c'est que la liaison départementale exhibée n'est pas une liaison temporelle. Naturellement le nombre de « liaisons départementales » éventuelles peut être fort élevé et il serait très long et fastidieux de les explorer toutes ; aussi aurons-nous recours à des *analyses de données* préalables pour déterminer a priori celles qui risquent d'être pertinentes, et écarter définitivement les autres (voir le paragraphe sur les variables exogènes).

3. — Choix des variables endogènes du modèle

Il s'agit de déterminer quelles caractéristiques statistiques de la criminalité répondent au mieux aux exigences suivantes :

a) être pertinentes pour s'introduire dans un modèle destiné à prévoir la population pénale ;

b) être différenciées suivant les départements, puisque la régression doit être effectuée sur les départements ;

c) évoluer de façon homogène dans le temps pour l'ensemble des départements, pour qu'il puisse y avoir coïncidence avec l'évolution temporelle ;

d) enfin, être des données statistiques dont la définition soit constante au cours du temps, pour éliminer des biais systématiques dus à la simple collecte des informations.

Compte tenu de la première contrainte, on pourrait imaginer d'utiliser pour chaque département le volume global des condamnations à des peines d'emprisonnement ferme ; mais une telle statistique ne pourrait pas évidemment être utilisée pour prévoir la population des PRÉVENUS. Utilisera-t-on alors le volume global par département de la criminalité légale qui, elle, existe et pourrait être reliée à la population des prévenus et des prisonniers condamnés. Malheureusement le chiffre de « criminalité légale » ne satisfait certainement pas à l'exigence (*d*) : il contient de nombreuses catégories d'infractions

« hétéroclites » dont la définition et l'évolution dans le temps pour certains départements n'a rien à voir avec l'évolution générale de la criminalité en France. C'est pourquoi on a été amené à extraire de la criminalité légale totale la part aberrante pour l'objectif à atteindre en opérant de la façon suivante : tout d'abord on a effectué (1) un regroupement des condamnations par catégories d'infractions correspondant à des types assez bien déterminés de comportements criminels — ce qui a conduit en particulier à placer dans une catégorie « condamnations diverses » les infractions hétérogènes qu'il était nécessaire d'éliminer. Finalement 7 catégories d'infractions ont été retenues pour classer les condamnations (2), avec les titres suivants :

- ASTUCIEUSES contre les BIENS
- VIOLENTES et BANALES contre les BIENS
- Contre la CHOSE PUBLIQUE
- INVOLONTAIRES contre les PERSONNES
- VOLONTAIRES contre les PERSONNES
- Contre les MŒURS
- aux règles de la CIRCULATION

Quoiqu'il en soit, après avoir défini ces « titres » de catégories d'infractions, il fallait encore déterminer le contenu qui satisfasse en particulier aux exigences (*b*), (*c*) et (*d*) ; le résultat a finalement été acquis après de nombreuses itérations. Citons quelques exemples de notre démarche à titre d'illustration : dans la catégorie « atteintes aux mœurs » on a omis le « racolage » dont la quasi-totalité est observée en 1968 à MARSEILLE (exigence *b*) ; de la catégorie « astucieuses contre les biens » on a extrait les « chèques sans provision » qui montrent une croissance explosive récente non décelable évidemment sur les départements à une date précédente comme 1962 (exigence *c*); de la catégorie « contre la chose publique » on a supprimé « chasse et pêche » qui y étaient comptées en 1968 alors qu'elles apparaissaient dans la catégorie « divers » en 1962 (exigence *d*), etc. Finalement le total des condamnations regroupées dans les 7 catégories d'infractions retenues représente environ 86 % de la criminalité légale (crimes, délits et contraventions de 5e classe) de la France en 1962 et 1968.

Arrivé à ce point on peut se demander si l'on ne pourrait pas justement caractériser chaque département par la répartition de sa criminalité entre les 7 catégories d'infractions, c'est-à-dire garder les 7 variables correspondantes comme variables endogènes dans le modèle liant la criminalité à la structure socio-professionnelle du département. Ceci peut être fort important, car on pourrait alors effectuer des projections pour chacune de ces catégories et donc, en particulier obtenir des renseignements intéressants sur les origines de la population pénale. Nous allons montrer que malheureusement cette procédure est IMPRATICABLE dans ce contexte, d'une part à cause de la nature

---

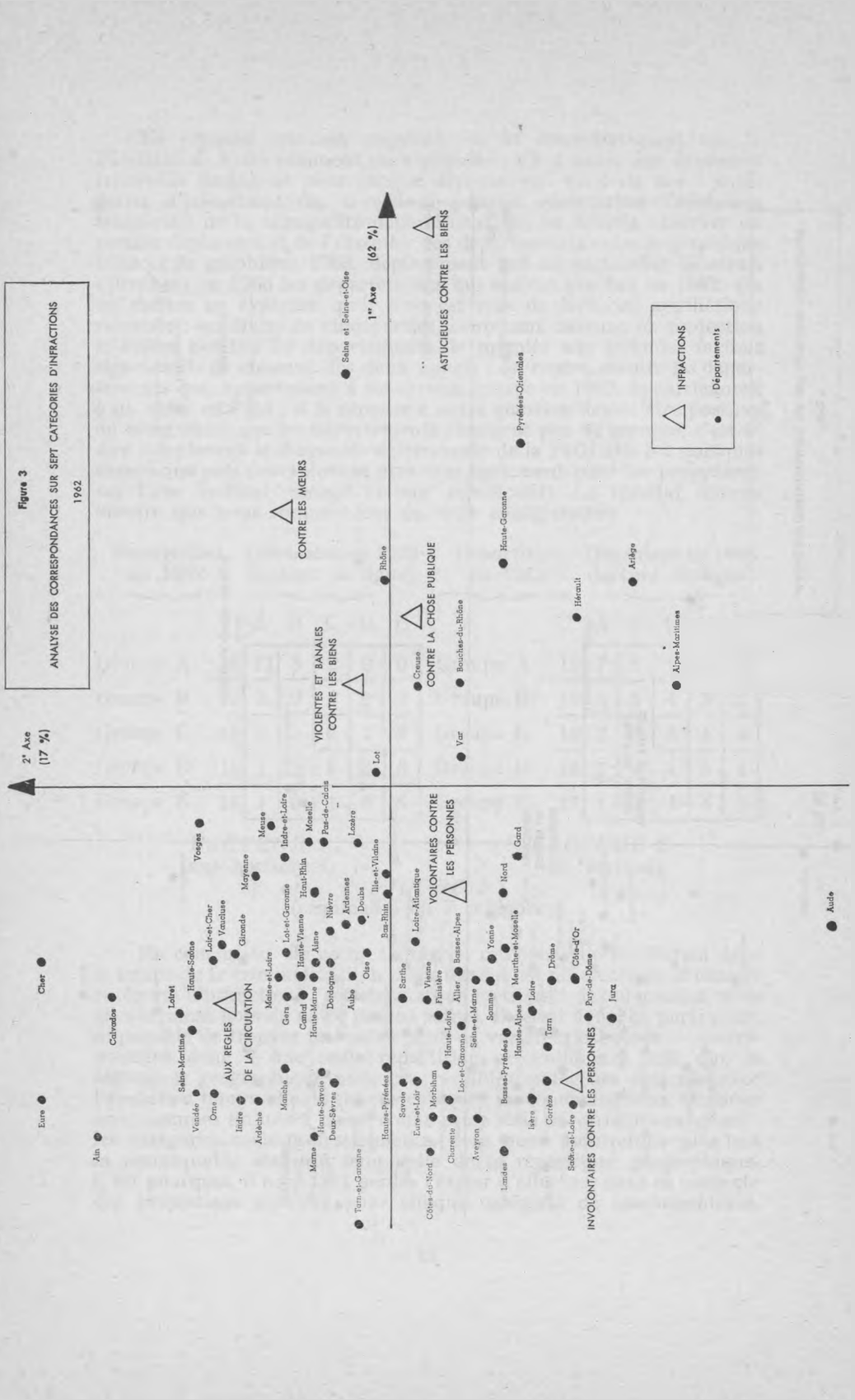
(1) Avec le concours du Service d'Etudes Pénales et Criminologiques et du Centre de Formation et de Recherche de l'Education Surveillée du Ministère de la Justice.

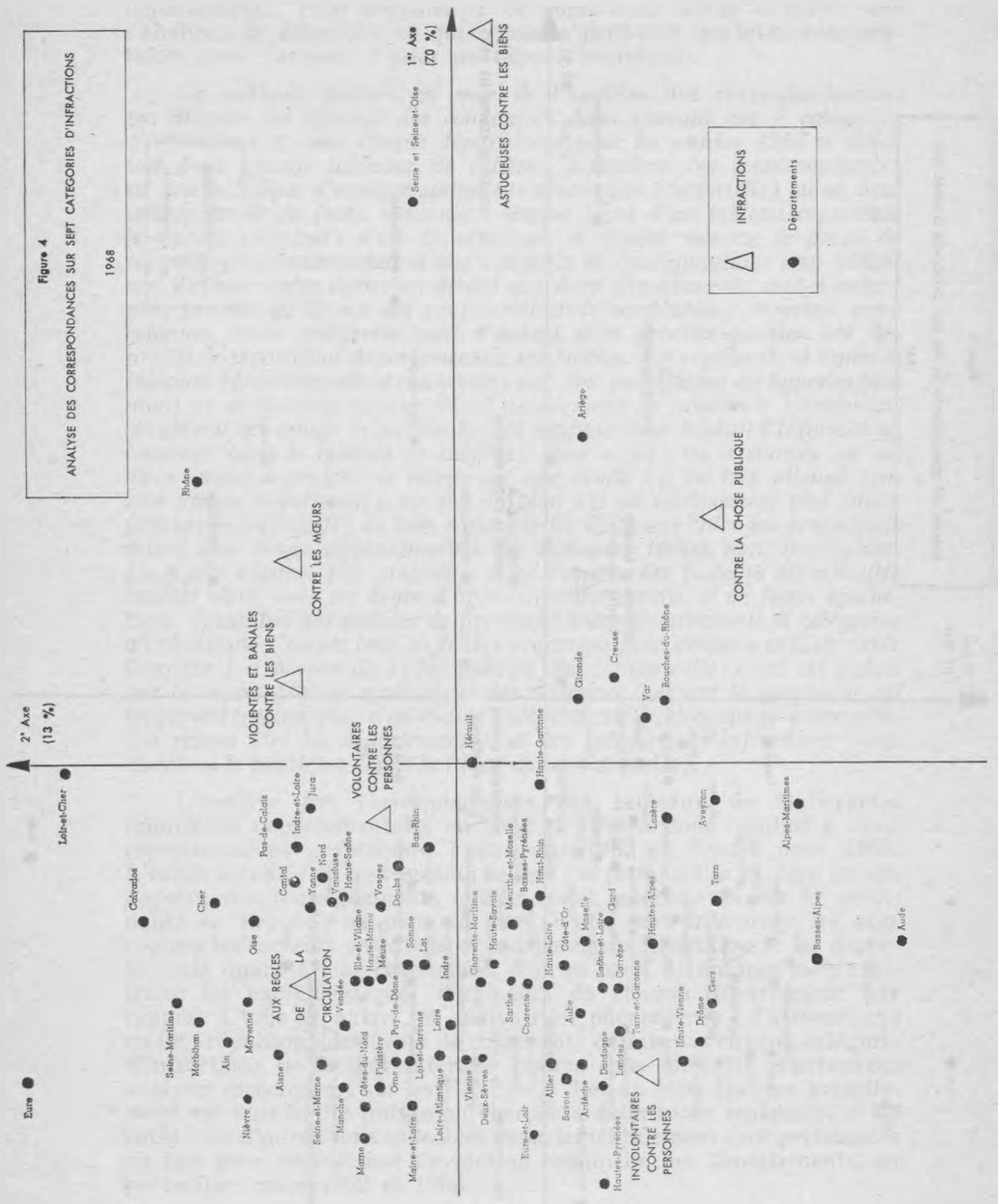
(2) On trouvera en annexe V la liste détaillée des infractions retenues dans chacune de ces catégories.

des 7 catégories retenues, d'autre part, parce qu'on a restreint le choix des variables exogènes aux seuls effectifs socio-professionnels des départements. Pour argumenter ce point nous allons effectuer des « analyses de données », ce qui nécessite peut-être quelques commentaires (voir l'annexe I pour un exposé technique).

La méthode utilisée est celle de l'analyse des correspondances. *On dispose des effectifs des condamnés dans chacune des 7 catégories d'infractions et pour chaque département pour les années 1962 et 1968, soit deux grands tableaux de chiffres. L'analyse des correspondances est une technique d'analyse statistique descriptive (factorielle) qu'on peut interpréter de la façon suivante : chaque ligne d'un tableau représente le « profil criminel » d'un département, et chaque colonne le profil de répartition départementale d'une catégorie de condamnations ; on définit une distance entre lignes en disant que deux départements sont d'autant plus proches qu'ils ont des profils criminels semblables ; de même entre colonnes, deux catégories sont d'autant plus proches qu'elles ont des profils de répartition départementale semblables. En représentant lignes et colonnes (départements et catégories) par des points dans un hyper-espace muni de la distance correspondant aux notions de proximité introduites, on obtient un nuage de points dont la configuration traduit l'information contenue dans le tableau de chiffres ; pour « voir » les distances on est alors amené à projeter ce nuage sur une droite s'il est très allongé (un seul facteur significatif), ou sur un plan s'il est relativement plat (deux facteurs significatifs) de telle sorte que les distances entre les projections soient une bonne approximation des distances réelles non observables. La figure obtenue par projection dans l'espace des facteurs significatifs traduit alors, avec un degré d'approximation connu, et de façon synthétique, l'essentiel des notions de proximité entre départements et catégories d'infractions. Compte tenu enfin des propriétés de la distance utilisée (voir l'annexe I à propos de « l'équivalence distributionnelle »), on est assuré que la représentation synthétique des distances lues sur le graphique est largement intrinsèque, c'est-à-dire indépendante du découpage géographique retenu (ici les départements), et des catégories d'infractions (voir ci-dessus le problème des définitions de ces catégories).*

L'analyse des correspondances des tableaux de statistiques criminelles départementales en 1962 et 1968 a donc conduit à deux représentations graphiques, l'une pour 1962 et l'autre pour 1968. L'étude détaillée de ces résultats ne sera pas retranscrite ici, bien qu'elle apporte des renseignements extrêmement intéressants sur la criminalité en France ; on peut en effet, d'une part *interpréter* les axes comme les facteurs essentiels et indépendants discriminant les départements quant à leur criminalité, d'autre part, déterminer sans arbitraire les caractéristiques criminelles de chaque département par rapport à tous les autres, et analyser les phénomènes « d'association » ou de « répulsion » des types de criminalité définis par chaque catégorie d'infractions (le lecteur intéressé pourra sans difficulté effectuer ces analyses directement sur les FIGURES 3 et 4). Notre propos actuellement est plus limité puisqu'on cherche à déterminer seulement si les catégories d'infractions collectées départementalement sont pertinentes ou non pour caractériser l'évolution criminelle des départements, en particulier entre 1962 et 1968.





La réponse, qui est négative, se lit immédiatement sur la FIGURE 5. Voici comment on a procédé : s'il y avait une évolution criminelle homogène pour chaque département vis-à-vis des 7 catégories d'infractions (ie. si celle-ci pouvait caractériser l'évolution temporelle de la criminalité départementale), on devrait observer un certain déplacement de *l'ensemble* des départements entre le graphique 1962 et le graphique 1968, déplacement qui en particulier laisserait « proches » en 1968 les départements qui étaient proches en 1962. On va mettre en évidence qu'il n'en est rien de la façon synthétique suivante : on divise en cinq parties, contenant chacune en projection le même nombre de départements, le premier axe (premier facteur significatif) de chacune des deux figures ; on repère ensuite les départements qui, appartenant à un certain groupe en 1962, appartiennent à un autre en 1968 ; si la réponse à notre question devait être positive on observerait que les départements changent peu de groupes, c'est-à-dire remplissent la diagonale descendante de la FIGURE 5 à quelques exceptions près ; ceci devant être vrai également pour les projections sur l'axe vertical (second facteur significatif). Le résultat obtenu montre que nous sommes loin de cette configuration.

Répartition en 1962		Dispersion en 1968 (lecture en ligne)				
		A	B	C	D	E
Groupe A	16	11	3	2	0	0
Groupe B	19	3	9	4	2	1
Groupe C	18	0	5	6	4	3
Groupe D	18	1	2	5	4	6
Groupe E	16	1	0	1	8	6

FACTEUR 1
(axe horizontal)

Répartition en 1962		Dispersion en 1968 (lecture en ligne)				
		A	B	C	D	E
Groupe A	17	7	6	1	2	0
Groupe B	19	5	5	4	3	2
Groupe C	18	2	3	6	4	3
Groupe D	18	1	4	4	5	4
Groupe E	17	1	0	4	4	8

FACTEUR 2
(axe vertical)

FIGURE 5
(criminalité en 7 catégories)

En conclusion de cette tentative, retenons que l'évolution dans le temps de la criminalité d'un département est différente pour chaque catégorie d'infractions ; cette remarque mériterait des précisions, mais elles seraient hors de notre propos actuel. Il serait donc, en particulier, impossible de trouver une *même série* de variables exogènes — correspondant donc à une seule répartition géographique) telle que la régression géographique avec ces variables endogènes coincide avec l'évolution temporelle de chaque catégorie de condamnations. Or, nous nous sommes limités a priori à une seule série de variables exogènes : les catégories socio-professionnelles, dont nous montrerons plus loin la remarquable stabilité temporelle de la répartition géographique. C'est pourquoi, il nous faut perdre l'espoir d'effectuer dans ce contexte des projections séparées pour chaque catégorie de condamnations.

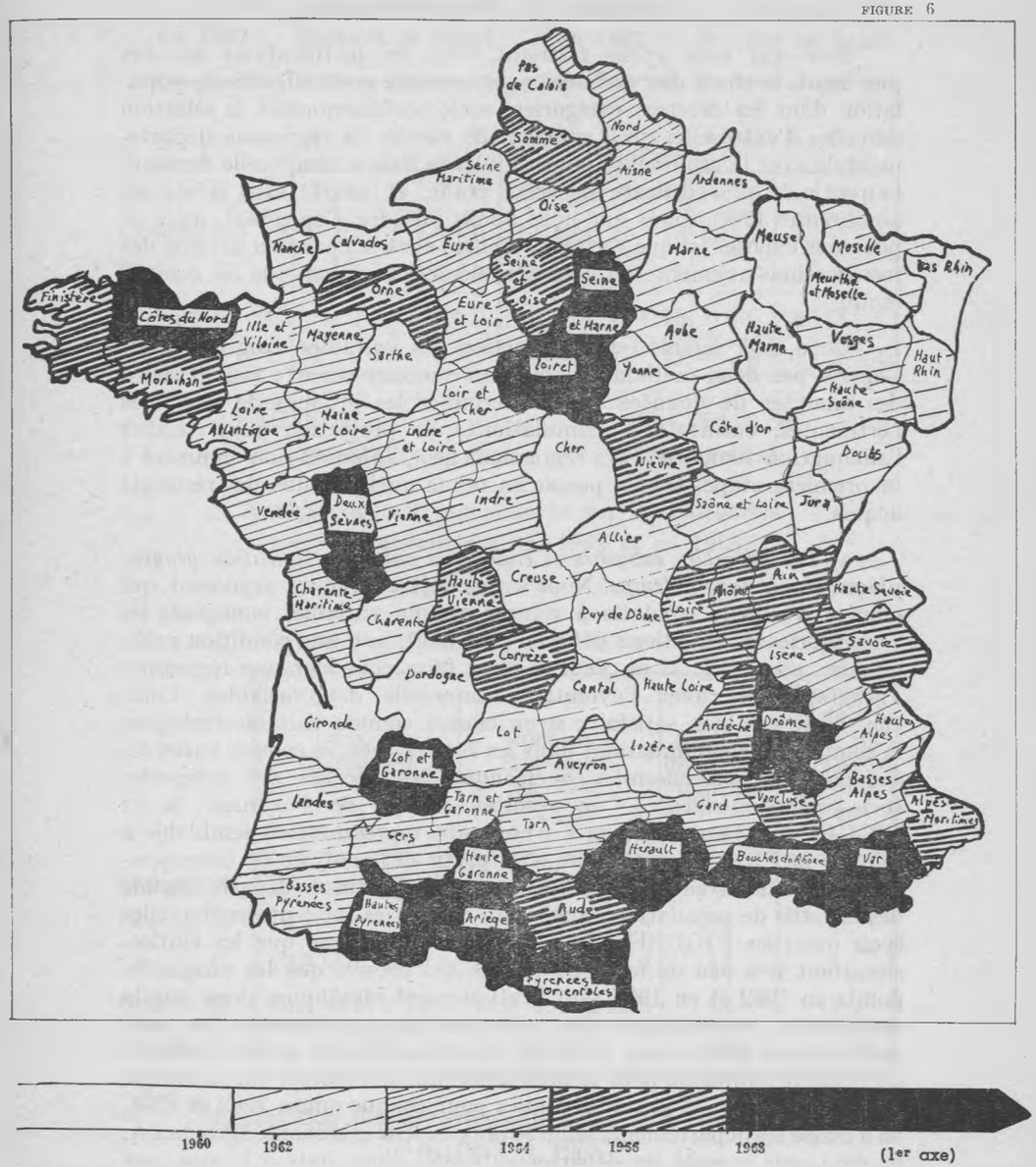
On peut songer alors à prendre comme caractéristique criminelle des départements, à la place de chacune des 7 catégories, *la SOMME des condamnations* correspondant à l'ensemble de ces 7 catégories d'infractions. Encore faut-il s'assurer qu'il s'agit bien là d'une donnée statistique qui différencie les départements quant à leur évolution dans le temps. La réponse, qui est positive, peut être lue directement sur le résultat d'une analyse de correspondances effectuée sur la série chronologique de ces sommes de condamnations relevées dans chaque département en 1960, 1962, 1964, 1966 et 1968. Il s'avère que seul le premier facteur extrait est significatif (le second étant une forme quadratique du premier) ; par conséquent, toute l'information contenue dans le tableau de chiffres est synthétisée par les proximités entre points-départements et points-années projetés sur cet axe. Or, il apparaît que les dates se projettent sur cet axe dans leur ordre chronologique et de façon très régulière, ce qui permet d'affirmer que le facteur extrait, donc le *seul* facteur expliquant les différences départementales de la somme des condamnations, est le FACTEUR TEMPS. Ce fait nous encourage donc à prendre comme caractéristique criminelle endogène dans le modèle la somme de ces condamnations. A titre indicatif, on a figuré les résultats de cette analyse par une représentation cartographique des départements sur la FIGURE 6 de la façon suivante : les départements qui se projettent à proximité d'une date sur l'axe extrait par l'analyse sont colorés d'une teinte uniforme qui va du blanc au noir quand on évolue de 1960 à 1968 ; autour d'une date donnée se retrouvent les départements qui ont subi à cette époque un plus grand accroissement de criminalité que les autres (l'examen de cette carte suggérera au lecteur des commentaires supplémentaires sur l'évolution de la criminalité).

Finalement, si on ajoute à ces arguments les justifications qui apparaîtront dans le paragraphe suivant (homogénéité temporelle et géographique semblable à celle des variables exogènes), on retiendra de cette discussion le point suivant :

La variable endogène retenue pour caractériser la criminalité et entrer en régression avec les indicateurs socio-professionnels en vue de simuler l'évolution temporelle de cette criminalité, sera le TOTAL des condamnations correspondant aux 7 catégories d'infractions définies plus haut.

Rappelons que ce total de condamnations représentait en 1962 et 1968 environ 86 % de la criminalité LEGALE de la France ; par conséquent les projections que nous allons effectuer pour 1975 correspondront vraisemblablement à un pourcentage approchant de la criminalité légale attendue à cette date. Ce qui pourrait être un inconvénient si le but à atteindre était effectivement des projections de la criminalité légale de la France, n'en est pas un pour nous qui ne recherchons qu'un intermédiaire de calcul adéquat pour obtenir des projections de population pénale.

FIGURE 6



ANALYSE DES CORRESPONDANCES

CRIMINALITE TOTALE DES ANNEES 1960, 1962, 1964, 1966, 1968

(1er axe : 57 % de la trace)

4. — Sélection des variables exogènes

Bien que nous ayons restreint, avec les justifications données plus haut, le choix des variables exogènes aux seuls effectifs de population dans les diverses catégories socio-professionnelles, la sélection de celles d'entre elles qu'il convient de mettre en régression départementale avec la criminalité pour simuler la liaison temporelle demeure la partie la plus délicate de cette étude, et mérite sans doute les nombreuses précautions que nous allons prendre. Cependant, nous ne pourrons exposer ici que certaines d'entre elles, un examen détaillé des justifications augmenterait considérablement le volume de ce compte rendu.

On a déjà signalé que la sélection des variables exogènes allait s'opérer par deux méthodes appliquées successivement ; tout d'abord des analyses de données pour déterminer les groupes de variables pertinentes, ensuite des « simulations » de projection pour vérifier l'adéquation temporelle des régressions géographiques non éliminées à la première étape. On va passer en revue quelques-uns des résultats acquis.

a) *Les variables exogènes et endogènes ont une répartition géographique stable dans le temps.* Nous avons déjà vu que cet argument, qui nous a d'ailleurs conduits à refuser comme variables endogènes les 7 catégories d'infractions définies plus haut, est une condition préliminaire essentielle si on veut espérer faire coïncider une régression « géographique » avec l'évolution temporelle des variables. Cette contrainte est bien satisfaite si on choisit comme variable endogène la somme des condamnations dans les 7 catégories, et comme variables exogènes un sous-ensemble quelconque des effectifs des catégories socio-professionnelles — ce résultat ressort de l'examen de la FIGURE 7. D'une part, on y trouve une représentation semblable à celle de la FIGURE 5 obtenue ici à partir des analyses de correspondance effectuées en 1962 et 1968 sur la répartition départementale des effectifs de population dans les 10 catégories socio-professionnelles (voir aussi les FIGURES 8 et 8 bis) ; on remarque que les chiffres s'écartent très peu de la diagonale, ce qui prouve que les nuages de points en 1962 et en 1968 sont pratiquement identiques, donc que la distribution géographique des catégories socio-professionnelles reste parfaitement stable dans le temps. Au-dessous on a mis en évidence un résultat semblable pour le taux de criminalité calculé sur la somme des infractions de la façon suivante : pour chaque année 1962 et 1968, on a classé les départements selon le taux de leur criminalité (groupes A, B, etc.) puis compté les départements qui, d'une date à l'autre, ont sauté plus d'une catégorie d'un classement à l'autre : l'examen des chiffres montre la stabilité temporelle du phénomène à un degré moindre cependant que pour les catégories socio-professionnelles.

Répartition en 1962		Dispersion en 1968 (lecture en ligne)				
		A	B	C	D	E
Groupe A	17	17	0	0	0	0
Groupe B	18	0	16	2	0	0
Groupe C	18	0	2	14	2	0
Groupe D	18	0	0	2	16	0
Groupe E	17	0	0	0	0	17

FACTEUR 1
(axe horizontal)

Répartition en 1962		Dispersion en 1968 (lecture en ligne)				
		A	B	C	D	E
Groupe A	17	11	6	0	0	0
Groupe B	18	6	7	5	0	0
Groupe C	18	0	5	10	3	0
Groupe D	18	0	0	2	13	3
Groupe E	17	0	0	0	3	14

FACTEUR 2
(axe vertical)

Analyse des correspondances sur les catégories socio-professionnelles en 1962 et en 1968.

	Taux de criminalité en 1962 (‰)	Taux de criminalité en 1968 (‰)	Nombre de départements	Nombre de départements déclassés de plus d'un groupe
Groupe A	$t \leqslant 10{,}8$	$t \leqslant 13{,}7$	17	2
Groupe B	$10{,}8 < t \leqslant 12{,}8$	$13{,}7 < t \leqslant 16{,}9$	18	1
Groupe C	$12{,}8 < t \leqslant 16{,}0$	$16{,}9 < t \leqslant 19{,}0$	17	5
Groupe D	$16{,}0 < t \leqslant 18{,}7$	$19{,}0 < t \leqslant 22{,}9$	18	2
Groupe E	$t > 18{,}7$	$t > 22{,}9$	18	6

Répartition des départements selon leur criminalité en 1962 et 1968

FIGURE 7

b) *Justification de l'emploi des variables.* Nous allons résumer dans ce paragraphe quelques-uns des arguments montrant qu'il y a bon espoir de trouver, avec les variables retenues, une régression géographique qui « simule » l'évolution temporelle des phénomènes mis en liaison. Considérons, par exemple, la FIGURE 8 qui est le résultat d'une analyse de correspondances sur les 10 catégories socioprofessionnelles en 1962 (il n'y a que deux facteurs significatifs). Cette représentation graphique présente deux particularités : d'une part, on a introduit, comme s'il s'agissait de simples départements, les unités statistiques « France entière » correspondant aux effectifs socio-professionnels globaux en 1962, en 1968 et en 1975 (projections de l'administration) ; d'autre part, on a classé les départements en groupe I, II, ..., VII suivant leur position sur le graphique. Comparons maintenant cette représentation à la FIGURE 8-bis qui correspond à la même analyse effectuée cette fois en 1968, et où les groupes contiennent les mêmes départements. Pour chaque groupe de départements, on a indiqué le taux moyen de sa criminalité aux dates où sont effectuées les analyses.

Sur les FIGURES 8 et 8-*bis* on peut remarquer tout d'abord la stabilité des groupes de départements, et ensuite une liaison très nette entre leur position et l'évolution du taux de criminalité moyen (qui n'intervient en aucun cas dans les calculs conduisant à ces représentations) : pour l'analyse 1962, en allant du groupe I au groupe VI, on trouve un taux de criminalité décroissant jusqu'au groupe II, puis croissant jusqu'au groupe VI en même temps qu'on suit l'évolution temporelle de la France 1962-1968-1975 ; si on se reporte maintenant à la véritable courbe d'évolution temporelle de la criminalité, on constate effectivement une telle évolution parallèle passant par un minimum autour des années 1955. Pour l'année 1968, en lisant dans le même sens du groupe I au groupe VI, on trouve des taux de criminalité toujours croissants et systématiquement plus élevés que dans la figure précédente. Tout se passe donc comme si on lisait une courbe parallèle à la courbe d'évolution temporelle de la criminalité en commençant dans ce cas un peu plus loin qu'on ne l'a fait pour la figure précédente, c'est-à-dire au-delà de la date où il y a un minimum de criminalité. D'autre part, les positions successives de la France entière 1962, 1968 et 1975 épousent encore la forme de cet axe d'évolution temporelle caractérisé par la position des groupes I à VI. On remarque enfin que, sur les deux figures, le groupe de départements n° VII se trouve hors de cet axe des temps. Il apparaît donc que l'analyse statistique à une date donnée de la répartition géographique des catégories socio-professionnelles (aux départements du groupe VII près) permet de reconstituer une certaine tranche de l'évolution temporelle du taux de criminalité. Cette constatation, faite en 1962 et en 1968, ne saurait être le fruit du hasard ; elle constituera en fait la justification majeure de nos calculs.

Pour rendre cette analyse plus facile à lire, on a remplacé sur les FIGURES 9 et 9-bis les départements par les *régions*, en prenant comme coordonnées des régions les barycentres des coordonnées des départements qui les composent (voir en annexe les propriétés de l'analyse pour justification). On peut voir sur ces figures que non seulement les *régions* s'ordonnent le long de l'axe temporel évoqué plus haut, parallèlement à la courbe joignant les points « France 62-68-75 », mais également les catégories socio-professionnelles, hormis la catégorie « OUVRIERS » qui, avec le groupe de départements n° VII et certaines régions, s'en écarte très significativement. Pour éviter d'allonger encore les commentaires sur cette adéquation de la répartition géographique des variables avec leur évolution temporelle, on a représenté sur la *FIGURE* 10 une carte synthétique où les régions sont classées selon leur position sur le premier facteur de l'analyse (qui coïncide donc avec la projection de « l'axe temporel ») et sont par ailleurs caractérisées par leurs taux de criminalité en 1962 et en1968.

L'étude de cette carte doit permettre au lecteur attentif de se convaincre qu'on puisse chercher à représenter la liaison temporelle entre criminalité et répartition socio-professionnelle par leur liaison géographique.

c) *Premières variables éliminées.* Parmi les 10 catégories socio-professionnelles étudiées globalement ci-dessus, certaines seront vraisemblablement mieux adaptées que d'autres pour satisfaire à notre exigence d'obtenir une liaison géographique avec la criminalité qui soit identique à la liaison temporelle. En particulier, il ressort de l'examen des FIGURES 9 et 9 bis que *la catégorie « OUVRIERS »*, qui se situe systématiquement hors de ce qu'on a appelé « l'axe temporel » traversant les régions (ou les départements), est sans conteste une variable qu'il ne faudra pas prendre en compte dans le modèle.

Pour des raisons que nous ne ferons qu'évoquer, il y a trois autres catégories qu'on peut éliminer *a priori*, à savoir :
les « NON-ACTIFS »
les « DIVERS »
les « PROFESSIONS LIBÉRALES ET CADRES SUPÉRIEURS »

Pour ce qui est des NON-ACTIFS et des DIVERS, ce sont des variables qui ont des répartitions géographiques très uniformes, comme on peut le constater par leur position centrale sur les FIGURES 9 et 9-bis et comme le confirme leur très faible écart type de répartition géographique ; ce sont donc des variables qui ne sont pas différenciées pour chaque département, et ont par conséquent peu d'intérêt pour nous. A cela s'ajoutent d'autres arguments ; signalons par exemple que si l'on effectue une analyse en *composantes principales* (voir l'annexe I) avec pour variables : les taux de criminalité départementale en 1962, 1964, 1966, 1968, les taux de « personnel de service » en 1962

FIGURE 8

ANALYSE DES CORRESPONDANCES

CATEGORIES SOCIO-PROFESSIONNELLES

— 1962 —

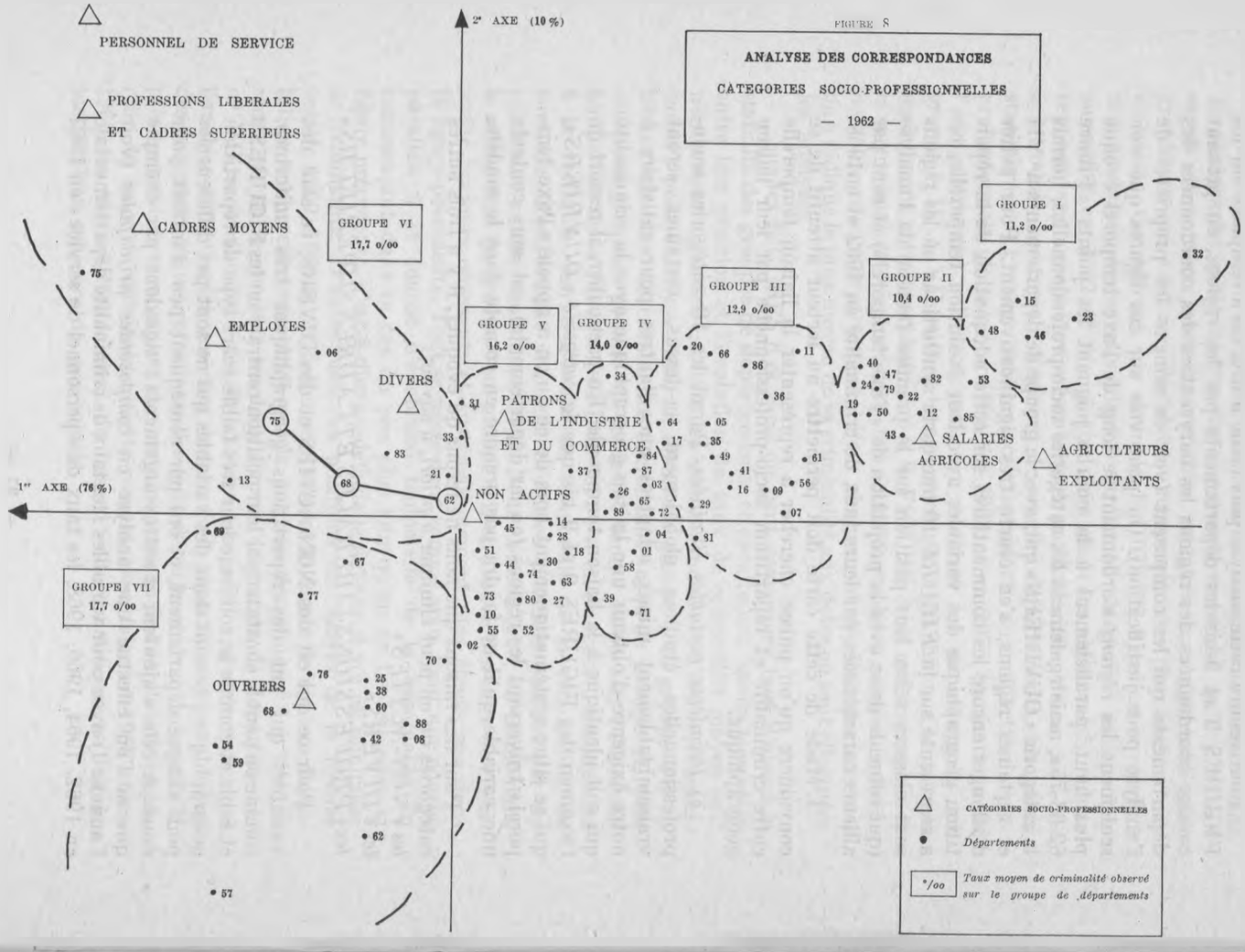


FIGURE 8 BIS

ANALYSE DES CORRESPONDANCES

CATEGORIES SOCIO-PROFESSIONNELLES

— 1968 —

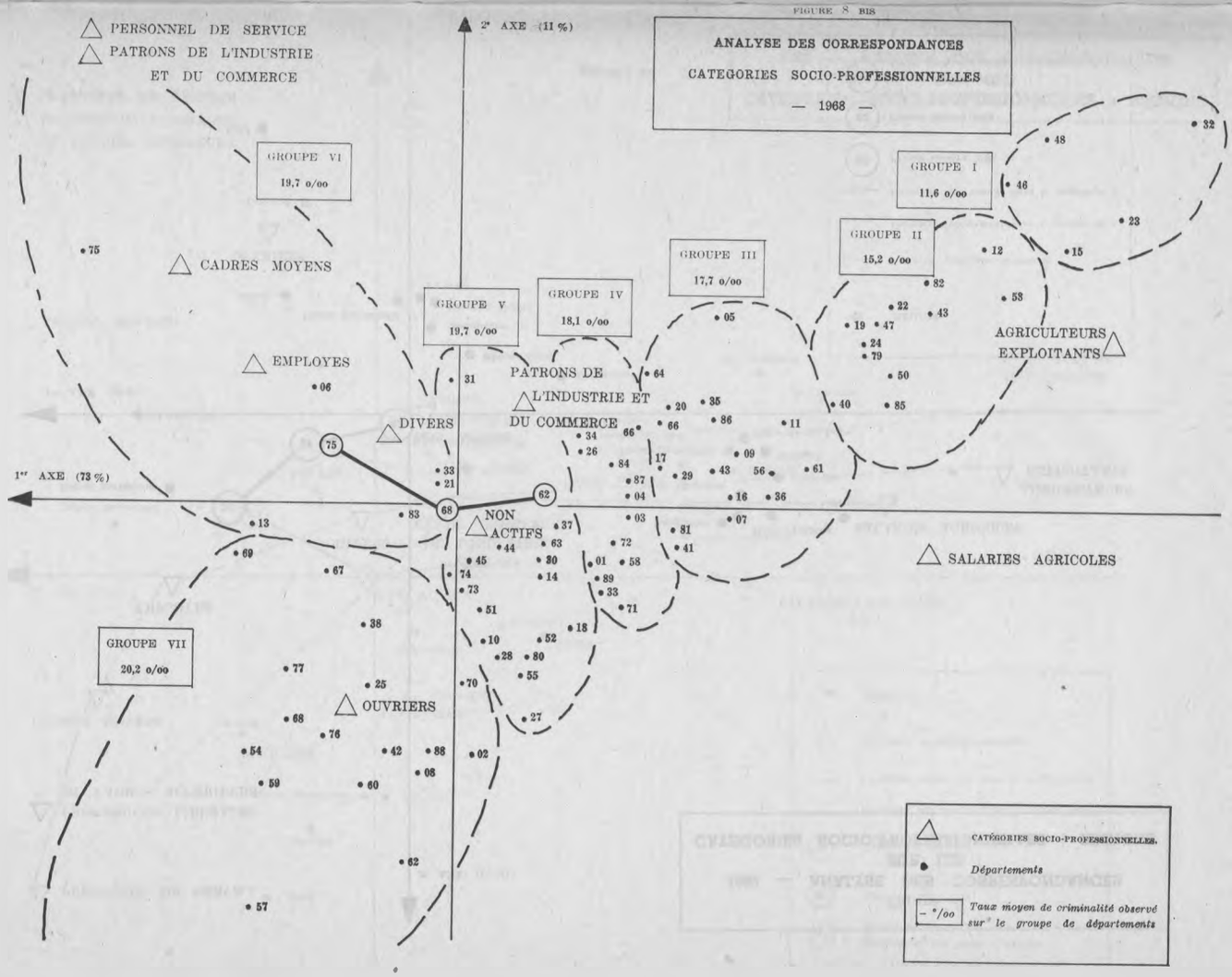


Figure 9

1962 — ANALYSE DES CORRESPONDANCES SUR LES CATEGORIES SOCIO-PROFESSIONNELLES : REGIONS

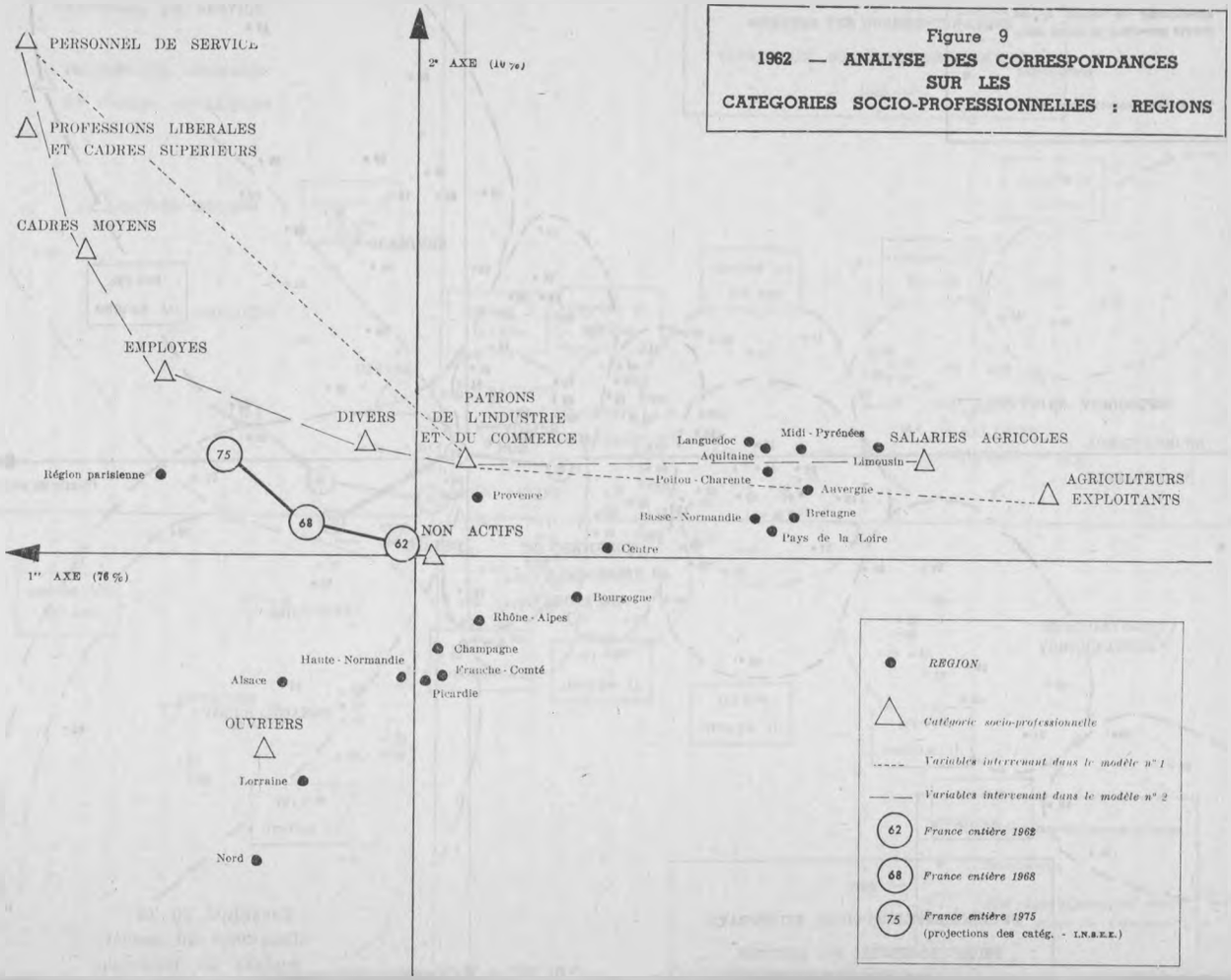


FIGURE 9 BIS

1968 — ANALYSE DES CORRESPONDANCES SUR LES CATEGORIES SOCIO-PROFESSIONNELLES : REGIONS

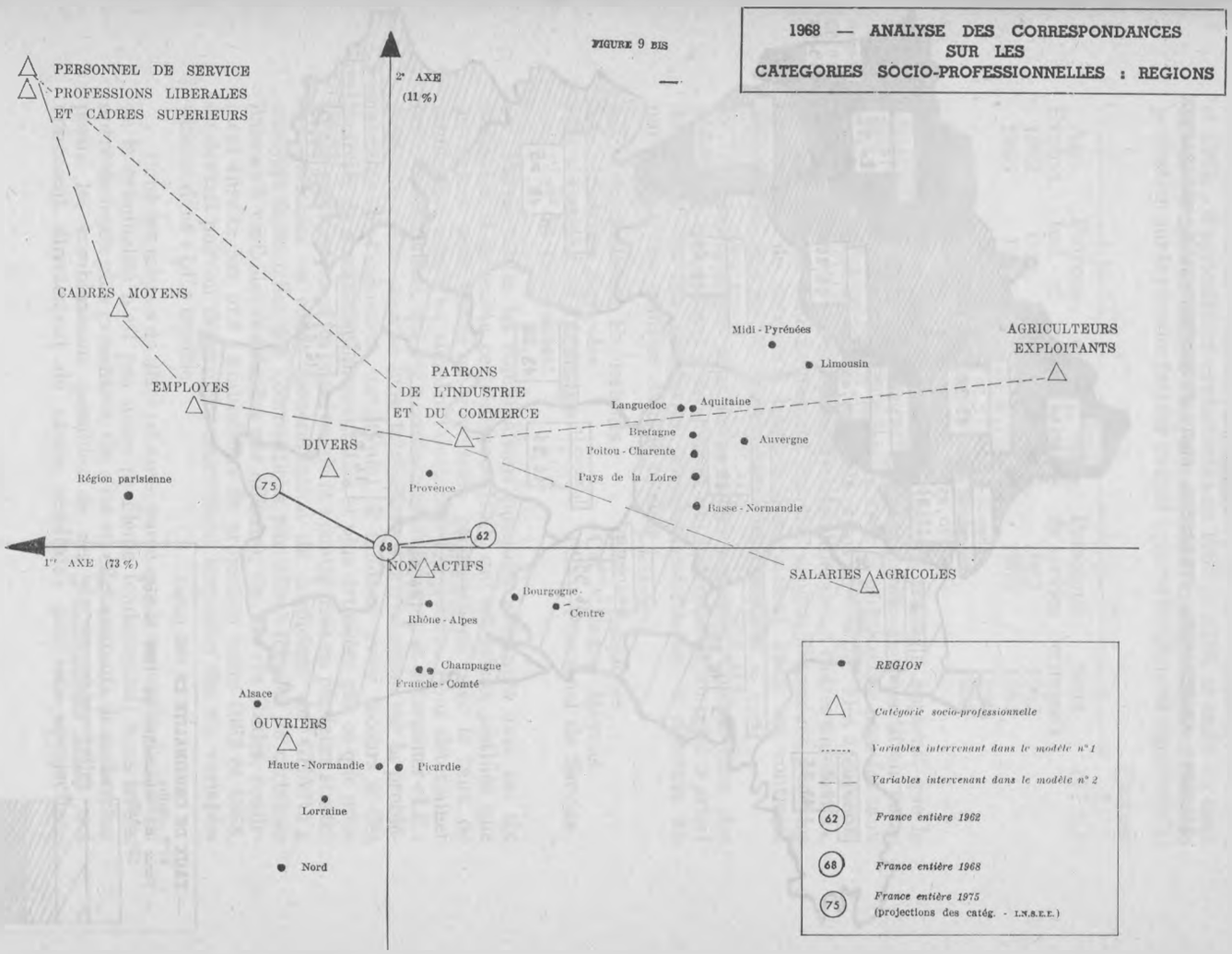
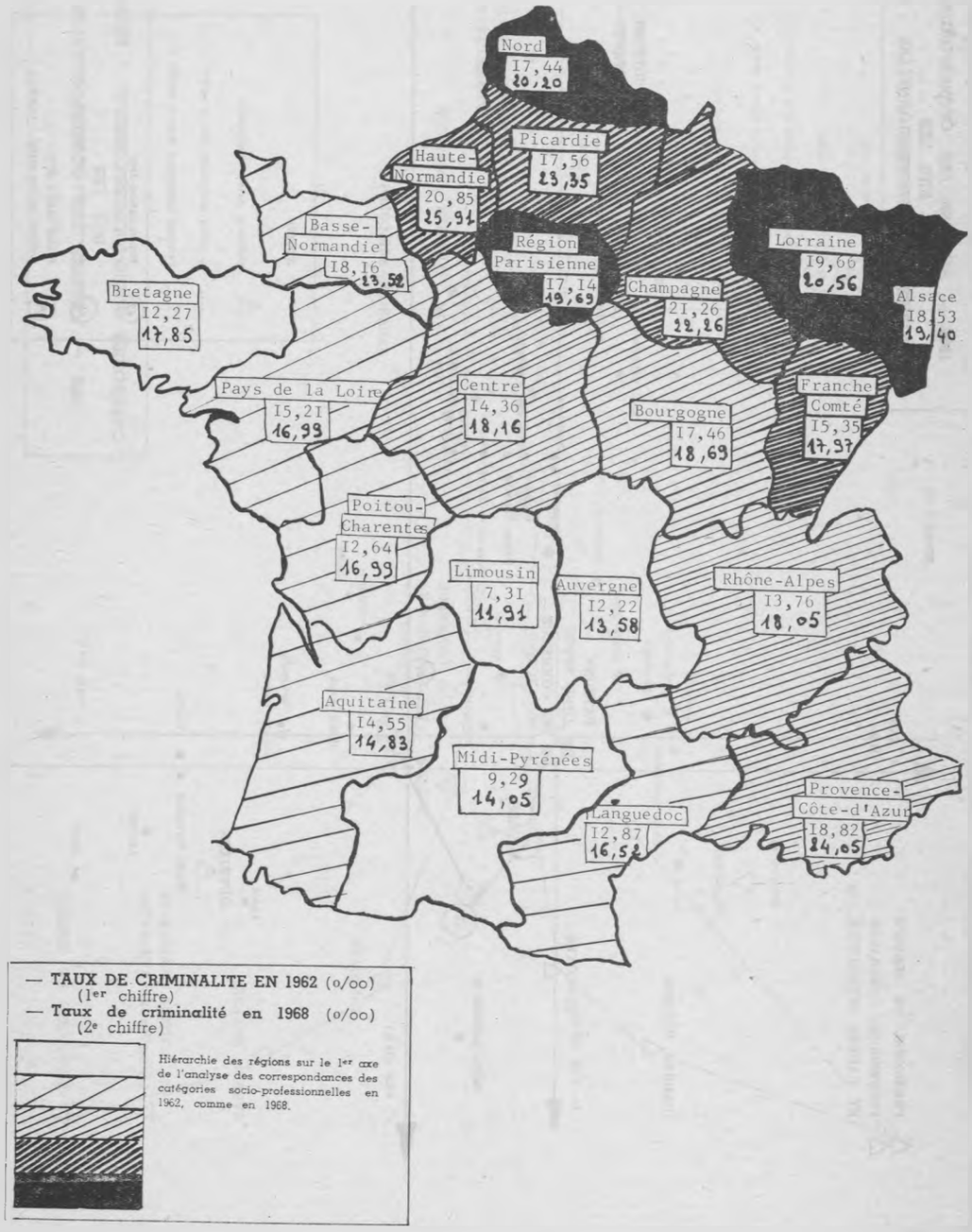
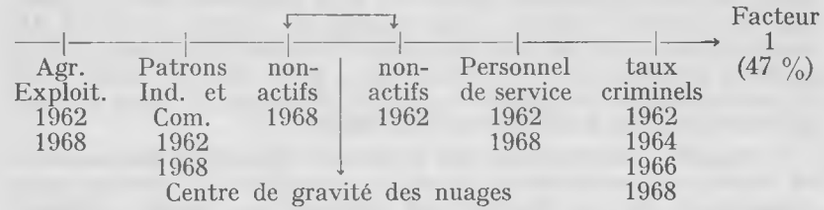


FIGURE 10

LIAISON : CRIMINALITE, STRUCTURE SOCIO-PROFESSIONNELLE, GEOGRAPHIE



et 1968, ceux de « patrons de l'industrie et du commerce » en 1962 et 1968, « d'agriculteurs exploitants » en 1962 et 1968, et enfin les taux de « non-actifs » en 1962 et 1968, on constate la configuration suivante en projection sur le premier facteur extrait (qui est hautement significatif) :



Sans entrer dans le détail de l'interprétation, il apparaît que la variable « non-actifs » présente une singularité de répartition entre 1962 et 1968 qui n'a rien de commun avec ce qui se passe pour les taux de criminalité et les autres variables prises en compte. Des arguments semblables tiennent pour les « professions libérales et cadres supérieurs », qui par ailleurs sont hautement correlés géographiquement avec les « personnels de services », lesquels apparaissent avoir au contraire de sérieuses raisons de figurer dans notre modèle.

d) *Problématique pour les variables restantes.* Compte tenu des arguments qu'on vient d'évoquer et qui ont conduit au rejet *a priori* de quatre des dix catégories socioprofessionnelles, il demeure en puissance d'être utilisées dans le modèle :

AE - Agriculteurs Exploitants	EM - Employés
SA - Salariés Agricoles	CM - Cadres Moyens
PIC - Patrons de l'Industrie et du Commerce	PS - Personnel de Service

Parmi toutes les combinaisons possibles effectuées avec ces six variables, il en existe certainement UNE qui est mieux justifiée que les autres pour entrer en régression géographique avec le taux de criminalité afin d'en simuler l'évolution temporelle. Pour déterminer cette combinaison, qui est nécessairement unique, et déterminer « LE » modèle *a priori* cherché (voir le chapitre d'introduction sur la problématique de l'induction statistique), il faudrait avoir recours à des analyses de données supplémentaires et plus détaillées que ce que nous avons pu faire jusqu'ici : analyse de la « contiguïté » en plusieurs strates des variables en cause (généralisation du coefficient de GEARY), analyses factorielles des « corrélations partielles » (études de certaines liaisons à variables constantes par ailleurs), etc. De ces études, confirmées ensuite par une « simulation » de projection entre 1962 et 1968, on devrait pouvoir déterminer exactement lesquelles des six variables utiliser dans « LE » modèle.

Pour des raisons de délai (et aussi parce que la recherche du modèle de la criminalité n'est pas notre problème fondamental) nous avons opéré de façon plus expéditive. On s'est en effet contenté de rechercher parmi les combinaisons possibles de ces six variables celles qui donnaient directement de « bons résultats » pour les « projections

simulées », c'est-à-dire représentaient apparemment la coïncidence entre liaison géographique et temporelle des phénomènes, sans chercher d'autres justifications. Evidemment cette procédure N'EST PAS rigoureuse, ni admissible en principe ; elle avait l'avantage cependant d'être très rapide, et notre propos est de la remplacer ultérieurement par la procédure correcte. Pour insister sur *l'aspect provisoire* des résultats obtenus ici, on a cru nécessaire de donner l'exemple de DEUX modèles construits de cette façon, rien *a priori* dans ce qu'on a fait ne permettant de discriminer le meilleur des deux, ni même s'il n'en existerait pas un troisième meilleur encore.

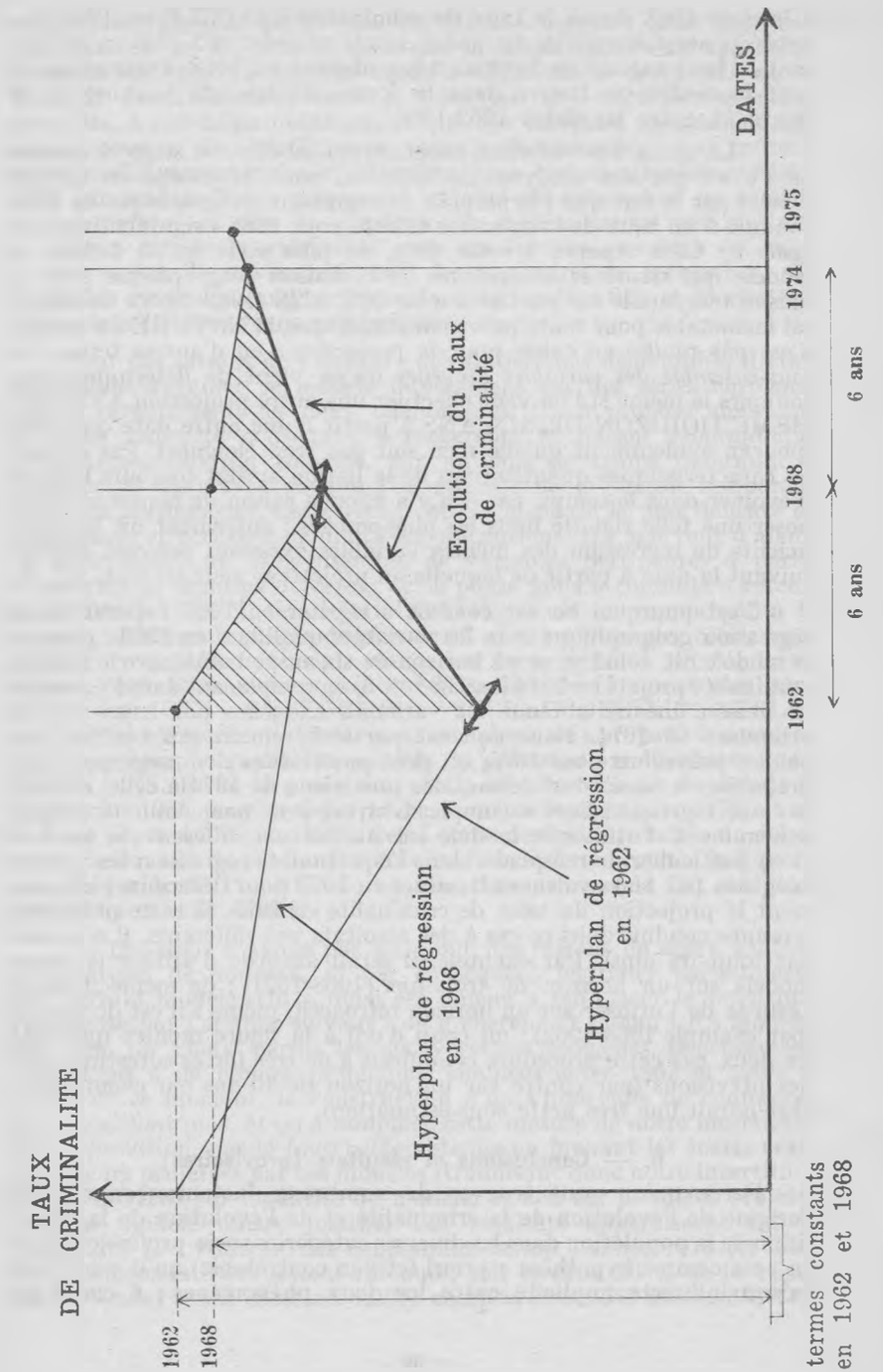
Signalons à ce propos un fait important. Chacun des deux modèles va conduire à des résultats différents, représentant en quelque sorte l'incertitude où on demeure sur l'identité du modèle véritable. Cependant la « fourchette » entre ces résultats ne représente en aucune manière les « fourchettes d'incertitude » déduites par induction statistique d'un modèle donné : une telle fourchette d'incertitude ne peut résulter en toute rigueur que des principes d'induction classique *provenant du terme aléatoire contenu dans le modèle*, alors que notre incertitude ici porte en fait sur l'écriture du modèle.

5. — Un modèle temporel à horizon déterminé

Il nous faut maintenant insister sur une caractéristique FONDAMENTALE du modèle que nous voulons construire, à savoir que ce modèle est défini (et donc utilisable) sur un *horizon temporel déterminé préalablement.* En d'autres termes, le modèle est construit pour représenter la coïncidence entre la liaison géographique d'une part et, d'autre part, la liaison temporelle des variables sur un certain laps de temps, et sur cette période seulement ; si un modèle est construit pour atteindre cette coïncidence sur une période de six ans par exemple, les variables exogènes choisies différeront a priori de celles d'un modèle destiné à simuler l'évolution temporelle sur un horizon de sept ans ou huit ans, etc. Ce fait est une conséquence inéluctable de la méthode, et l'oublier pourrait conduire à de graves erreurs comme nous allons l'évoquer.

La FIGURE 11 ci-dessous va nous permettre de « visualiser » le raisonnement, bien qu'en toute rigueur il faudrait la tracer dans un espace ayant au moins quatre dimensions. L'axe horizontal est l'axe des temps, alors que l'axe vertical portera les taux de criminalité de la France. Une régression géographique à une date déterminée, par exemple 1962, définit un hyperplan de régression figuré sur le dessin par une double flèche à la verticale de la date étudiée ; le terme constant dans l'équation de régression se trouve représenté alors par l'intersection de l'hyperplan avec l'axe vertical (c'est-à-dire ici dans le prolongement de la double flèche). Comment étudie-t-on par exemple l'adéquation de la formule de régression pour simuler l'évolution de la criminalité en 1968? On est conduit à remplacer dans l'équation de l'hyperplan de 1962 les variables exogènes par les valeurs qu'elles prendront en 1968. Graphiquement ceci se traduit de la façon suivante : sur la verticale de la date 1962 on reporte le terme constant de l'équation et on déplace parallèlement à lui-même l'hyperplan de régression pour qu'il passe par ce point ; son intersection avec la verticale élevée

FIGURE 11



à la date 1968 donne le taux de criminalité CALCULÉ en 1968. On apprécie alors *la validité du modèle sur la période* 1962-1968 en comparant le taux calculé en 1968 au taux observé en 1968. Pour résumer, tout le modèle se trouve dans la forme du triangle hachuré de la figure 11 entre les dates 1962-1968.

Le choix des variables exogènes du modèle est suggéré comme on l'a vu par des analyses de données, mais déterminé en dernier ressort par le fait que l'hyperplan de régression géographique en 1962 conduit à un taux de criminalité calculé pour 1968 « significativement égal » au taux observé à cette date, de telle sorte qu'on définit un modèle qui assure la coïncidence de la liaison géographique avec la liaison temporelle sur un horizon de SIX ANS (1962-1968). Comme il est inéluctable pour toute prévision statistique, la NATURE du modèle n'est pas remise en cause pour la projection ; en d'autres termes, *le sous-ensemble des variables exogènes* qu'on vient de déterminer sera toujours le même si l'on veut effectuer une autre projection AVEC LE MEME HORIZON DE SIX ANS à partir d'une autre date que 1962 (pourvu évidemment qu'elle n'en soit pas trop éloignée). Par contre, les caractéristiques quantitatives de la liaison auront toujours le droit d'évoluer dans le temps, car il n'y a aucune raison de penser et d'imposer une telle rigidité dans les phénomènes ; autrement dit les coefficients de régression des mêmes variables exogènes peuvent évoluer suivant la date à partir de laquelle on projettera pour six ans.

C'est pourquoi on est conduit à répéter en 1968 l'opération de régression géographique avec les variables qualifiées en 1962 ; puisque le modèle est défini pour un horizon de six ans, on obtiendra le taux de criminalité projeté en 1974 (ie. 1968 + 6) en remplaçant dans l'équation de liaison linéaire obtenue les variables exogènes par leurs valeurs attendues en 1974. Puisqu'on est particulièrement intéressé en fait par les prévisions pour 1975, on peut passer sans dommage pour une projection à aussi court terme, des prévisions de 1974 à celles de 1975 par une régression linéaire simple. L'erreur à ne pas commettre serait évidemment d'utiliser le modèle sur un horizon différent de six ans, et en particulier de remplacer dans l'équation de régression les valeurs exogènes par leurs valeurs attendues en 1975 pour déterminer directement la projection du taux de criminalité en 1975. Si cette procédure erronnée conduit dans ce cas à des résultats peu différents, il n'en sera pas toujours ainsi. Par exemple, il serait absurde d'utiliser le même modèle sur un horizon de trois ans (1968-1971) ; de même il serait absurde de l'utiliser sur un horizon rétroactif, même s'il est de six ans (par exemple 1968-1962) ; un coup d'œil à la figure montre que dans les deux cas cette procédure conduirait à de très fortes surestimations des prévisions (par contre sur un horizon de 10 ans par exemple, on obtiendrait une très nette sous-estimation).

6. — **Conclusions et résultats (provisoires)**

Parce qu'on peut trouver de nombreux facteurs communs à l'origine de l'évolution de la criminalité et de l'évolution de la répartition de la population dans les diverses catégories socio-professionnelles on pose comme hypothèse a priori (et non contrôlable) qu'il existe une liaison indirecte implicite entre les deux phénomènes ; à cause du

champ relativement étroit des variations observées pour les deux phénomènes, on suppose de plus que la liaison peut être approchée correctement par un modèle linéaire. Signalons à ce propos qu'on pourrait éventuellement enrichir l'éventail des variables exogènes possibles, à condition évidemment que ces variables supplémentaires aient, comme les statistiques socio-professionnelles, un contenu social suffisamment riche et complexe pour être éventuellement liées à la criminalité. Notre propos est en effet de déterminer un modèle de simulation, et non de construire le modèle causal qui mettrait en liaison directe la criminalité avec ses diverses « causes » (et dans ce cas certes il conviendrait tout au contraire de rechercher des variables exogènes « explicatives » dont la signification et le contenu soient très étroits et sans ambiguïté).

Pour des raisons de « qualité » des statistiques disponibles, il était impossible d'étudier ce modèle directement sur des séries chronologiques caractérisant les deux phénomènes. On a proposé alors une méthode indirecte consistant à rechercher la liaison géographique (départementale) qui coïncide, sur un laps de temps déterminé à l'avance (6 ans) avec ce que l'on connaît de la liaison temporelle entre les phénomènes. Cette recherche est la phase la plus délicate de la méthode et nécessite de multiples précautions. Il s'agit tout d'abord de sélectionner un sous-ensemble de variables pertinentes à partir d'analyses de données diverses ; cette phase nous a conduits à retenir d'un côté le TAUX de CRIMINALITÉ calculé en rapportant à la population masculine âgée de plus de 18 ans la somme des condamnations correspondant à sept catégories bien définies d'infractions, d'autre part, le taux de population masculine de 15 ans et plus dans SIX des DIX CATÉGORIES SOCIO-PROFESSIONNELLES (I.N.S.E.E.). Ensuite doit intervenir une phase de « projection simulée » pour apprécier la qualité du modèle de prévision sur l'horizon choisi de SIX ANS. Enfin la projection réelle est effectuée en réajustant le modèle en 1968 pour projeter (avec le même horizon) en 1974, et en déduire la prévision pour 1975. Cette étape devrait être suivie du calcul d'un « intervalle de confiance » autour de la valeur projetée indiquant en quelque sorte l'incertitude où l'on demeure par le fait des aléas introduits dans le modèle.

En fait nous n'avons pas pu, pour des raisons de délai, achever les analyses de données, ni par conséquent légitimer parfaitement UN certain modèle. On a donc été amené à remplacer la procédure rigoureuse par l'artifice suivant : on a cherché et retenu LES modèles qui, par un choix parmi les six variables exogènes sélectionnées, conduisent à de bonnes projections simulées entre 1962 et 1968. A partir de ce moment, la construction d'un « intervalle de confiance » ne se justifiait plus, et on a remplacé cette mesure de notre incertitude sur la prévision, par la fourchette obtenue en prenant les écarts entre les valeurs projetées par ces modèles (traduisant donc notre incertitude PROVISOIRE sur l'identité du modèle, et non sur la valeur projetée par un modèle mieux justifié). Les résultats apparaissent sur la *FIGURE 12* : l'effectif des condamnés durant l'année 1975 (et dans les 7 catégories d'infractions retenues, soit environ 86 % de la criminalité totale) serait donc un chiffre compris entre 358 631 et 381 117.

FIGURE 12 — *PROJECTIONS DE LA CRIMINALITÉ (1) EN* 1975

VARIABLES (2)		1974
Y	taux de criminalité	
SA	taux de Salariés Agricoles	2,00
AE	taux d'Agriculteurs Exploitants	6,60
PIC	taux de Patrons de l'Industrie et du Commerce	6,65
CM	taux de Cadres Moyens	8,10
EM	taux d'Employés	7,00
PS	taux de Personnels de Service	1,38

MODÈLE N° 1

Équation 1962 ..... Y = 1,7138 (PS) — 0,9333 (PIC) — 0,2143 (AE) + 24,4537
Écarts-types ..... (0,98) (0,40) (0,07)
Projection simulée 1968 Y = 18,45 (écart à la vraie valeur : 2,5 %)

Équation 1968 ..... Y = 2,8448 (PS) — 0,4399 (PIC) — 0,2357 (AE) + 21,7758
Écarts-types ..... (1,32) (0,45) (0,09)
Projection 1974 ..... Y = 21,22 ‰

MODÈLE N° 2

Équation 1962 ..... Y = 2,8077(PS) + 1,2209(EM) — 0,6921(CM) — 1,4870(PIC) — 0,0065(SA) + 21,5446
Écarts-types ..... (1,14) (0,71) (0,76) (0,42) (0,15)
Projection simulée 1968 Y = 18,24 (écart à la vraie valeur : 3,7 %)

Équation 1968 ..... Y = 5,6682(PS) + 0,6688(EM) — 0,6311(CM) — 1,3396(PIC) + 0,2743(SA) + 21,2168
Écarts-types ..... (1,49) (0,81) (0,70) (0,49) (0,23)
Projections 1974 .. Y = 20,12 ‰)

Projections

	1962	1968	Criminalité projetée	
			en 1974	en 1975
Taux de criminalité (‰)	15,84	18,94	*modèle 1* 21,220 *modèle 2* 20,122	21,521 20,251
Population (en milliers)	15 580,56	16 896,60	17 548,80	17 708,80
Criminalité	246 788	320 021	*modèle 1* 372 389,0 *modèle 2* 353 115,2	381 117,0 358 630,9

(1) Rappelons que l'effectif criminel considéré représente en fait 86 % environ de la criminalité totale en 1962 et 1968.

(2) On trouvera en annexe VI, pour 1962 et 1968, les séries départementales de criminalité et de répartition par catégorie socio-professionnelles sur lesquelles ont été exécutés les calculs.

III. — Seconde étape

PROJECTION DES PRÉVENUS

1. — Introduction

Un problème fondamental apparaît en ce qui concerne la prévision de la population des prévenus lorsque l'on sait que de nouvelles ou récentes mesures législatives sont destinées justement à modifier la structure et l'évolution de cette catégorie de présents dans les prisons. Dans ces circonstances, il eût été maladroit d'utiliser une méthode de projection « systématique » comme celle exposée au chapitre précédant, c'est-à-dire fondée sur un modèle rigide sur lequel le Décideur (ie. l'Administration) n'aurait eu aucun contrôle ; au contraire il semble particulièrement opportun de mettre au point dans ce cas précis un modèle qui permette d'étudier l'impact éventuel de mesures législatives, c'est-à-dire un modèle explicitant des « variables instrumentales » dont les valeurs puissent être modulées par le Décideur.

Il va de soi qu'une telle entreprise ne saurait être simple ; et notre propos n'est ici que de faire le premier pas dans cette voie. On signalera, en particulier, au cours de ce chapitre, divers points traités ici de façon trop simplifiée et sur lesquels on est décidé à revenir ultérieurement si on veut assurer une crédibilité plus grande aux modèles.

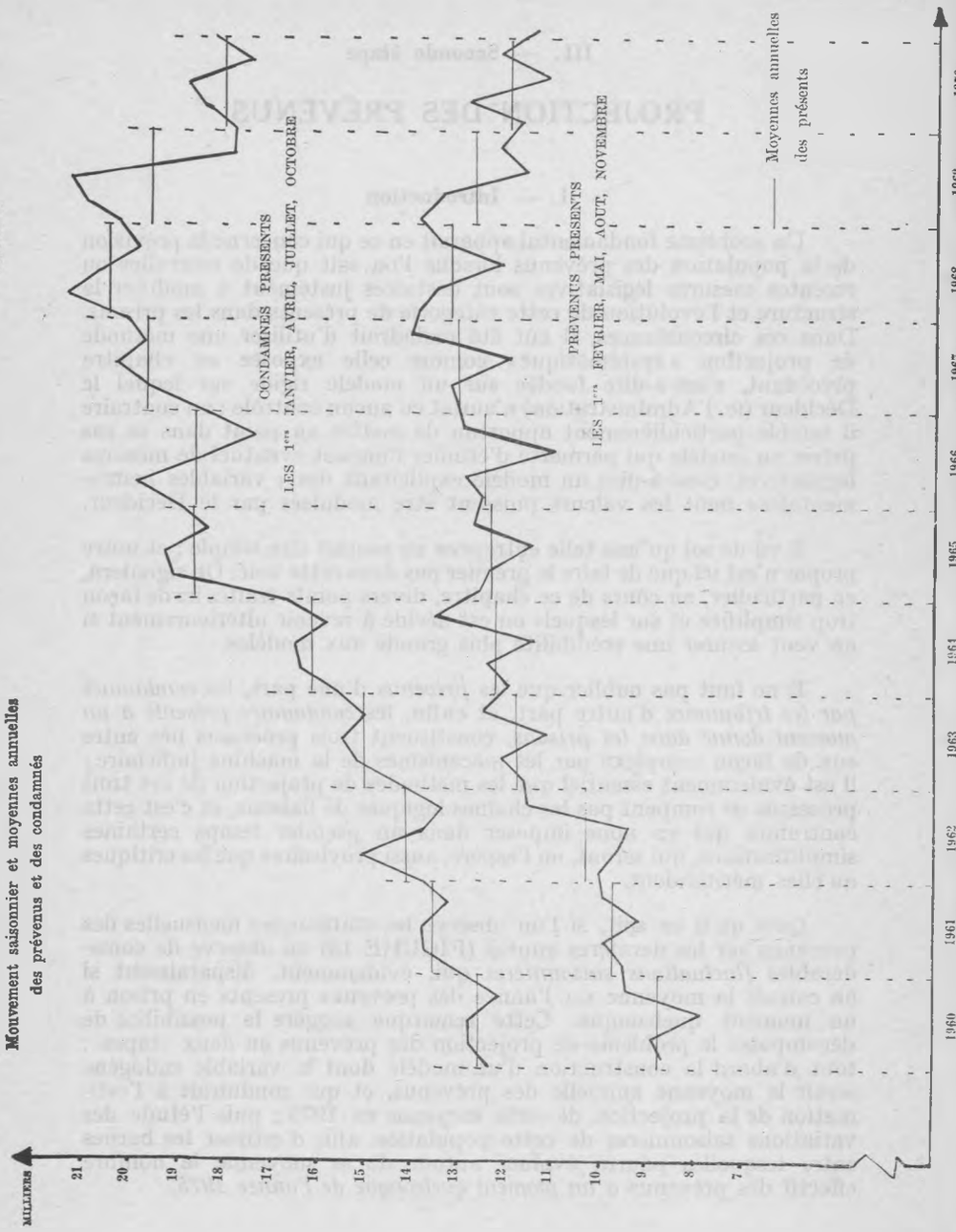
Il ne faut pas oublier que les *prévenus* d'une part, les *condamnés par les tribunaux* d'autre part, et enfin, les *condamnés présents à un moment donné dans les prisons*, constituent trois processus liés entre eux de façon complexe par les mécanismes de la machine judiciaire ; il est évidemment essentiel que les méthodes de projection de ces trois processus ne rompent pas les chaînes logiques de liaisons, et c'est cette contrainte qui va nous imposer dans un premier temps certaines simplifications, qui seront, on l'espère, aussi provisoires que les critiques qu'elles mériteraient.

Quoi qu'il en soit, si l'on observe les statistiques mensuelles des prévenus sur les dernières années (FIGURE 13) on observe de considérables *fluctuations saisonnières* qui, évidemment, disparaissent si on calcule la moyenne sur l'année des prévenus présents en prison à un moment quelconque. Cette remarque suggère la possibilité de décomposer le problème de projection des prévenus en deux étapes : tout d'abord la construction d'un modèle dont la variable endogène serait la moyenne annuelle des prévenus, et qui conduirait à l'estimation de la projection de cette moyenne en 1975 ; puis l'étude des variations saisonnières de cette population afin d'estimer les bornes entre lesquelles pourra évoluer autour de sa moyenne le nombre effectif des prévenus *à un moment quelconque de l'année 1975.*

FIGURE 13

POPULATION PENALE

Mouvement saisonnier et moyennes annuelles des prévenus et des condamnés



2. — Le modèle

Compte tenu de ce qui vient d'être dit, il faut s'attendre à ce que les hypothèses retenues pour définir le modèle a priori pour la population des prévenus, seront de nature tout à fait différente des hypothèses retenues pour le modèle « criminalité ». Tout d'abord, il est évidemment exclu de pouvoir étudier le phénomène de façon départementale car il est impossible de ventiler géographiquement la population pénale ; d'autre part, dans la mesure où l'on étudiera des phénomènes liés DIRECTEMENT entre eux, il ne peut guère s'élever d'objection à ce qu'on observe directement leur évolution temporelle simultanée. C'est pourquoi nous allons chercher si l'on peut justifier une liaison a priori entre le processus temporel des effectifs moyens de prévenus et celui des effectifs de condamnés par les tribunaux. Cette liaison va apparaître en fait comme résultante logique de trois hypothèses. Nous évoquerons plus loin comment ces hypothèses, qui peuvent apparaître trop simplificatrices, sont en fait largement justifiées par leurs conséquences, et pourront de toute façon être diversifiées par la suite.

PREMIÈRE HYPOTHÈSE : le bilan des entrées-sorties de prison comptabilisées du premier janvier jusqu'à la date t de l'année n est une fonction linéaire croissante du nombre de PRÉVENUS entrant en prison à la date t de l'année n.

L'interprétation de cette hypothèse a priori est simple et, semble-t-il, naturelle : il reste d'autant plus d'individus en prison sur une période donnée qu'il entre davantage de nouveaux prévenus — plus précisément : qu'il entre davantage de nouveaux prévenus à la fin de la période. Cette précision, qui peut sembler anodine, est en fait l'essence de l'hypothèse, et on va voir qu'elle est destinée à traduire une certaine INERTIE ou stationnarité du processus d'entrées-sorties des prisons.

L'indice n repère l'année, la variable t indiquera une certaine date au cours de l'année, tandis que dt représentera un petit intervalle de temps. Appelons $P_n(t)$ la fonction qui donne pour chaque date t de l'année n le nombre de prévenus présents en prison à cette date ; durant l'intervalle de temps dt, le nombre de prévenus présents aura varié de la quantité :

$P_n(t) - P_n(t + dt)$ soit $dP(t)$ (dérivée de la fonction). Cette variation est due à l'arrivée de nouveaux individus et à la sortie d'un certain nombre de prisonniers pendant la durée dt ; appelons $E_n(t)$ [resp. $S_n(t)$] la fonction de répartition donnant le nombre d'entrées (resp. de sorties) à la date t de l'année n. Il existe alors une relation logique inéluctable entre le processus des entrées, le processus des sorties et le nombre de présents à un moment donné :

$$d\,P_n(t) = E_n(t)\,dt - S_n(t)\,dt$$

soit w une variable muette ; cette équation s'écrit aussi bien après intégration :

$$(1)\quad P_n(t) = P_n(o) + \int_o^t [E_n(w) - S_n(w)]\,dw$$

où $P_n(o)$ désigne le nombre de prévenus présents au 1er janvier ($t = o$) de l'année n. L'équation (1) est une relation objective où n'entre aucune hypothèse a priori.

Traduisons maintenant de façon formelle l'hypothèse encadrée plus haut. En appelant toujours $E_n(t)$ le nombre de nouveaux prévenus entrant à la date t, et en désignant par a et b des coefficients inconnus, elle s'écrit :

(2) $P_n(o) + \int_o^t [E_n(w) - S_n(w)]\, dw = a\, E_n(t) + b$

compte tenu de l'équation (1), on peut donc écrire :

(3) $P_n(t) = a\, E_n(t) + b$

Nous allons intégrer l'équation (3) pour t décrivant l'année n (ie. : $n \leqslant t < n+1$), avec $\int_n^{n+1} dt = 1$:

(4) $\int_n^{n+1} P_n(t)\, dt = a \int_n^{n+1} E_n(t)\, dt + b$

Cette relation est facile à interpréter : dans le membre de gauche on trouve la MOYENNE $\overline{P}_n$ au cours de l'année n du nombre de prévenus présents ; quant à l'intégrale du membre de droite, elle représente la somme E_n de toutes les nouvelles entrées de l'année n. En d'autres termes, l'équation (4) s'écrit :

(5) $\overline{P}_n = a\, E_n + b$

On aperçoit sur la formule (5), équivalente à l'hypothèse énoncée, comment cette hypothèse traduit en fait une certaine stationnarité du phénomène d'entrées-sorties de prisons : quelques soient les fluctuations saisonnières des nouveaux entrants au cours de l'année, les sorties évoluent aussi et *de telle sorte que la moyenne des prévenus présents au cours de l'année ne soit fonction que du total des nouveaux entrants de l'année.*

SECONDE HYPOTHÈSE : Si tous les prévenus entrés au cours de l'année n devaient être jugés avant que s'écoule un an, on observerait que le nombre des condamnés résultant serait proportionnel — à un résidu aléatoire près — au nombre des entrants.

Cette hypothèse concerne évidemment le mode de fonctionnement de la machine judiciaire. On sait qu'un certain pourcentage, sans doute faible mais mal connu, de prévenus subissent une durée de prévention supérieure à un an ; l'hypothèse stipule que, si ce pourcentage était réduit à zéro, c'est-à-dire si tous les prévenus étaient jugés avant que s'écoule un an de prévention, il en résulterait un certain nombre de condamnations au cours de cette année là, proportionnel au nombre des affaires jugées (à des fluctuations aléatoires près).

Nous allons traduire cette hypothèse de façon formelle. Appelons toujours E_n le nombre total des entrées au cours de l'année n, et soit C^*_{n+1} le nombre (fictif) des condamnations au cours de l'année suivante

si l'hypothèse est satisfaite ; soit c un coefficient inconnu (coefficient de proportionnalité) et v_n une variable aléatoire de moyenne nulle (fluctuation aléatoire pour l'année n) ; alors :

$C^*_{n+1} = c\, E_n + v_n$

Il va de soi que le coefficient de proportionnalité c de cette hypothèse doit être positif. A ce propos, nous nous apercevrons plus loin que, pour atteindre le modèle de projection cherché, on peut se passer dans une première étape de l'estimation de ce coefficient inconnu. Par contre, dans une phase ultérieure de la recherche, où on va tenter d'exhiber des « variables instrumentales » pour le Décideur, il sera bien sûr nécessaire d'étudier de façon précise cet ajustement.

TROISIÈME HYPOTHÈSE : A des fluctuations aléatoires près, l'accroissement des condamnations entre les années n et $n+1$ est proportionnel à l'ÉCART entre l'effectif qu'il faudrait condamner durant l'année $n+1$ si la prévention ne devait pas durer plus d'un an, et l'effectif des condamnés durant l'année n.

Cette hypothèse est très importante ; elle est destinée, comme on va le voir, à traduire explicitement un certain processus d'AUTO-RÉGULATION de la machine judiciaire. Avant d'en préciser l'interprétation et la justification naturelle, on peut la transcrire de façon formelle. Appelons d un coefficient inconnu (coefficient de proportionnalité) et soit w_n une variable aléatoire de moyenne nulle (fluctuation aléatoire pour l'année n) ; alors la troisième hypothèse s'écrit :

$C_{n+1} - C_n = d\,(C^*_{n+1} - C_n) + w_n$

Durant l'année n, on a prononcé C_n condamnations ; on a observé simultanément l'arrivée de E_n nouveaux entrants. Si tous ces individus devaient être jugés dans l'année qui suit leur entrée, on devrait effectuer C^*_{n+1} condamnations durant l'année $n+1$ (d'après la seconde hypothèse) ; en fait, la troisième hypothèse dit qu'on prononcera un nombre réel C_{n+1} de condamnations certainement SUPÉRIEUR à C^*_{n+1}. On peut interpréter cette hypothèse comme une réaction d'autorégulation de la machine judiciaire pour limiter les « effets de saturation » si les entrées E_t de l'année n sont telles qu'on devrait en toute rigueur traiter dans l'année $n+1$ un nombre d'affaires C^*_{n+1} supérieur au nombre d'affaires traitées effectivement dans l'année n, alors on aura tendance à traiter en « urgence », et pour éviter l'accumulation des retards, des affaires en attente depuis un certain temps, de sorte que le bilan effectif C_{n+1} de l'année $n+1$ sera supérieur à ce qu'on aurait dû attendre C^*_{n+1}. Evidemment l'hypothèse laisse place à des fluctuations aléatoires représentées par la variable aléatoire w_n.

Notons que l'interprétation de cette hypothèse stipule que le coefficient de proportionnalité soit supérieur à 1, pour que C_{n+1} soit effectivement supérieur (aux aléas près) à C^*_{n+1} ; cette condition devra naturellement être satisfaite par l'estimateur du coefficient c lorsqu'on opérera l'induction statistique sur le modèle. Signalons enfin que, compte tenu du faible pourcentage d'individus passant plus d'un an en détention préventive, la quantité C^*_{n+1} ne devrait être que *très*

légèrement inférieure à C_{n+1} ; d'ailleurs l'espérance mathématique de la quantité $(C_{n+1} - C^*_{n+1})\ C_{n+1}$ pourra être utilisée comme estimateur du pourcentage de prévenus dont la durée de prévention a dépassé une année.

3. — Autre formulation du modèle

Nous allons tout d'abord résumer brièvement cette présentation du modèle sous son aspect formel ; il est caractérisé par trois relations, dont les deux dernières font intervenir (comme le modèle linéaire) des variables aléatoires justifiant les méthodes d'induction statistique :

$$\text{(i)}\quad \bar{P}_n = a\,E_n + b \qquad \text{avec } a > 0$$

$$\text{(ii)}\quad C^*_{n+1} = c\,E_n + v_n \qquad \text{avec } c > 0$$

$$\text{(iii)}\quad C_{n+1} - C_n = d\,(C^*_{n+1} - C_n) + w_n \qquad \text{avec } d \geqslant 1$$

Sous cette forme il s'agit donc d'un modèle du type « à équations multiples » et pour lequel, comme on sait, l'induction statistique directe risque d'être fort délicate en raison des problèmes de sur-identification ; d'autre part, ce modèle a la particularité de faire intervenir une variable non-observable directement : C^*_{n+1}. On est donc amené à transformer le modèle, sans cependant modifier les hypothèses, mais uniquement pour l'écrire sous une forme à laquelle puisse s'appliquer l'induction statistique classique. Cette transformation va consister à éliminer entre les trois équations les variables qui ne sont pas déterminantes : E_n et C^*_{n+1}.

De l'équation (i), on tire :

$$E_n = \frac{1}{a}\bar{P}_n - \frac{b}{a}$$

en portant cette expression dans l'équation (ii), il vient :

$$C^*_{n+1} = \frac{c}{a}\bar{P}_n - \frac{bc}{a} + v_n$$

on peut écrire alors dans l'équation (iii) :

$$C_{n+1} - C_n = \frac{cd}{a}\bar{P}_n - \frac{bcd}{a} + dv_n - dC_n + w_n$$

équation qui peut encore s'écrire :

$$\text{(iv)}\quad \begin{cases} \bar{P}_n = a\dfrac{(d-1)}{cd}C_n + \dfrac{a}{cd}C_{n+1} + b - \left(\dfrac{a}{c}v_n + \dfrac{a}{cd}w_n\right) \\ \text{avec } a > 0,\ c > 0,\ d \geqslant 1 \end{cases}$$

Sous cette forme, qui est une conséquence logique des trois hypothèses énoncées plus haut pour définir le modèle, on voit que les effectifs des condamnations qui seront prononcées durant l'année n et durant l'année $n + 1$ sont reliées intimement au nombre moyen de prévenus présents en prison au cours de l'année n. Cette conséquence, qui trouve une interprétation naturelle, peut servir à confirmer a posteriori les hypothèses que nous avons choisies ; on sait, en effet, que la majorité des prévenus passe moins d'un an en détention préventive, de sorte qu'en régime stationnaire, ils détermineront une certaine proportion des condamnés de l'année n et une part moindre des condamnés de l'année $n + 1$. Par conséquent on aurait pu écrire directement une équation de la forme :

$$\text{(v)}\quad \bar{P}_n = x_0\,C_n + x_1\,C_{n+1} + x_2 + u_n$$

où x_0, x_1 et x_2 sont des coefficients inconnus et u_n un résidu aléatoire de moyenne nulle. Or l'expression (v) est formellement identique à l'expression (iv) écrite plus haut ; il suffit en effet de poser :

$$\begin{cases} x_0 = \dfrac{a}{cd}(d-1) \ ; \ x_1 = \dfrac{a}{cd} \ ; \ x_2 = b \\ u_n = -\left(\dfrac{a}{c}v_n + \dfrac{a}{cd}w_n\right) \end{cases}$$

v_n et w_n ayant des espérances mathématiques nulles, il en sera de même de u_n ; de plus l'hypothèse a priori $d \geqslant 1$ entraîne effectivement $x_0 \geqslant x_1$ correspondant au fait qu'une majorité de prévenus sont jugés dans l'année de leur incarcération.

Pourquoi dans ces conditions n'avoir pas posé directement l'équation (v) comme modèle a priori, plutôt que s'astreindre à définir un modèle à partir d'hypothèses de comportement telles les trois hypothèses citées, et finalement aboutir à la même équation? On peut tout d'abord arguer qu'il est plus satisfaisant pour l'esprit de savoir ce qui est à l'origine de ce qu'on observe ; il n'est sans doute pas dénué d'intérêt d'avoir « décortiqué » certains éléments du mécanisme de fonctionnement judiciaire (processus d'entrées-sorties, processus d'autorégulation, etc.) et d'avoir montré comment la combinaison complexe de ces mécanismes se traduit finalement par la conséquence vérifiable aisément que les prévenus « alimentent » les condamnés de l'année, et en portion moindre ceux de l'année suivante. Enfin, ce passage par les « hypothèses de comportement » était évidemment la porte étroite à franchir si l'on veut ultérieurement dégager dans ces processus les « variables instrumentales » sur lesquelles le Décideur peut agir par des mesures législatives pour infléchir les évolutions ou la structure de la population des prévenus.

Nous allons pouvoir maintenant aborder le problème de l'induction statistique puisque dans l'équation (iv) (ou dans l'équation (v) formellement identique) il n'intervient que des variables observées pour lesquelles on dispose de séries chronologiques annuelles.

4. — Induction statistique sur le modèle

Sous la forme de l'équation (iv) ou de l'équation (v) on se trouve en présence d'un modèle du type « à retards échelonnés », où le nombre de « lags » à prendre en compte sur la variable exogène est parfaitement connu : C_n et C_{n+1}. Autrement dit ce modèle ne présente pas la difficulté classique des modèles à « retards échelonnés », où on demande à l'induction statistique une estimation simultanée des coefficients du modèle et du nombre de « retards » significatifs. Grâce à cette particularité, rien ne s'oppose à considérer ce modèle comme un modèle linéaire ordinaire, et à effectuer la régression multiple correspondante pour en estimer les coefficients.

Cependant si la méthode est parfaitement justifiée d'un point de vue théorique, on sait que sa mise en œuvre possède elle aussi des caractéristiques théoriques particulières. En particulier il faut s'attendre à ce que la *variance* des estimateurs des coefficients soit relativement large ; ce phénomène est expliqué par la multicollinéarité inévitable entre les séries $\{C_n\}$ et $\{C_{n+1}\}$, du fait de l'évolution assez régulière de la variable au cours du temps (voir l'annexe technique sur le modèle linéaire ; la matrice de variance-covariance des estimateurs s'obtient par inversion d'une matrice « proche » d'une matrice singulière). On aura soin par conséquent de vérifier sur les calculs que les écarts-types calculés pour les coefficients de C_n et C_{n+1} ne sont pas exagérément grands, faute de quoi l'ajustement risquerait d'être très imprécis.

On sait par ailleurs que si l'ajustement est effectué directement sur le modèle « à retards échelonnés », le corrélogramme calculé sur les résidus de l'ajustement fournit en général une estimation *sans biais* du corrélogramme théorique des erreurs. Cette propriété autorise donc à tester une éventuelle liaison temporelle des aléas (u_n) de l'équation du modèle à l'aide du test classique de DURBIN-WATSON (1950); il est important, en effet, de vérifier qu'il n'y a pas autocorrélation significative des (u_n) si l'on veut être sûr que les estimations des coefficients de C_n et C_{n+1} par l'ajustement linéaire sont SANS BIAIS.

La prévision de la population des prévenus en 1975 pourra alors être effectuée selon les principes généraux, en procédant de façon itérative à partir de l'estimation de l'équation du modèle, et en utilisant comme série des variables exogènes C_n, C_{n+1} les calculs de projection effectués au chapitre précédent. Sous réserve des hypothèses classiques, cette méthode minimise en effet le risque d'erreur sur la projection. De même pourra-t-on calculer un *intervalle de confiance* pour cette projection de prévenus, en utilisant les formules données dans l'annexe technique. Rappelons à cette occasion qu'en étudiant le modèle de la criminalité (chapitre précédent), l'inachèvement des analyses nous a conduits à conserver provisoirement deux modèles alternatifs entre lesquels on n'a pas tranché ; ceci nous a donc fourni en fait deux résultats de projections criminelles pour 1975. De ce fait, l'utilisation de l'un et l'autre résultats dans le modèle de projection des prévenus va nous conduire également à deux calculs de projection pour la moyenne annuelle des prévenus, calculs que nous repérerons par un numéro : nº 1 correspondant au modèle 1 de la criminalité, et nº 2 correspondant

au modèle 2. Précisons encore une fois que l'existence de ces deux résultats ne saurait être que provisoire, et doit rappeler au lecteur l'état d'inachèvement de cette étude.

5. — Les résultats

Les données utilisées sont présentées dans le tableau de la FIGURE 14, et les résultats sur la figure suivante. On trouvera une représentation graphique de l'évolution sur la FIGURE 16.

FIGURE 14

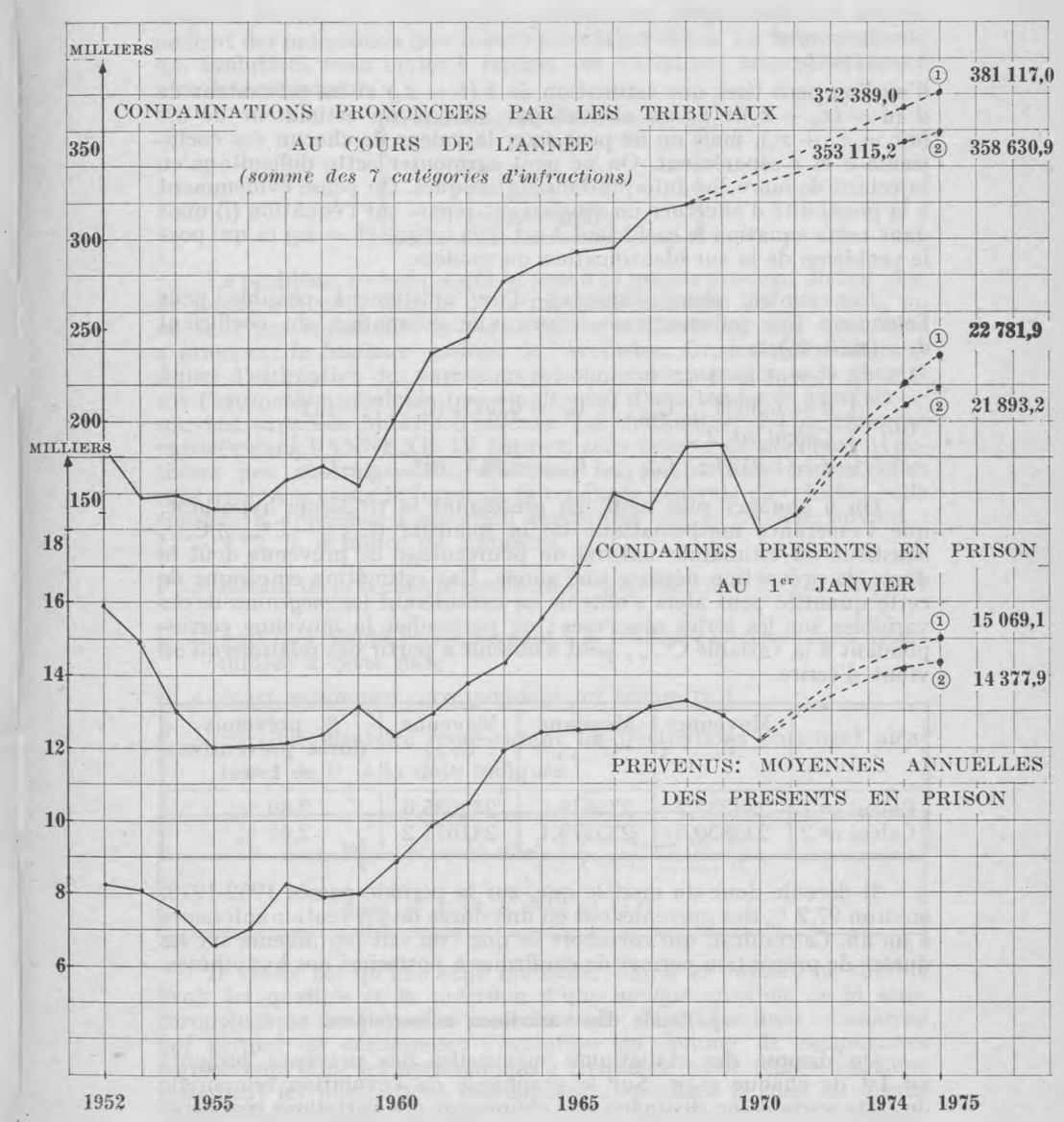
Date	$\overline{P_n}$ (*)	C_n (Modèle 1)	C_n (Modèle 2)	
1952	8 286	187 595		
1953	8 066	157 878		
1954	7 558	159 504		
1955	6 565	152 120		
1956	6 997	152 120		
1957	8 260	167 989		
1958	7 900	170 649		
1959	7 956	164 627		
1960	8 824	223 837		OBSERVATIONS
1961	9 826	238 717		
1962	10 475	246 800		
1963	11 927	260 901		
1964	12 447	277 286		
1965	12 506	293 172		
1966	12 587	295 897		
1967	13 172	317 077		
1968	13 354	320 021		
1969	12 919	328 749	325 537	
1970	12 205	337 477	331 052	PROJECTIONS
.....		.....	.....	
.....		.....	.....	
1975		381 117	358 631	

(*) Ces moyennes annuelles ont été calculées sur les effectifs des prévenus de droit commun (hommes et femmes), présents en prison au 1er jour de chaque mois. (Source : Bureau de la détention.)

FIGURE 15

CALCUL n° 1	Equations	$\bar{P}_n = 0{,}02854\ C_n + 0{,}00532\ C_{n+1} + 2\,115{,}7$
	Ecarts-types	(0,00932) (0,00886)
	Projections 1975	$\bar{P}_{75} = 15069{,}1$
	Intervalle de confiance à 95 %	$14340{,}0 \leqslant \bar{P}_{75} \leqslant 15798{,}2$
CALCUL n° 2	Equations	$\bar{P}_n = 0{,}02849\ C_n + 0{,}00592\ C_{n+1} + 2\,002{,}3$
	Écarts-types	(0,00868) (0,00832)
	Projections 1975	$\bar{P}_{75} = 14377{,}9$
	Intervalle de confiance à 95 %	$13765{,}2 \leqslant \bar{P}_{75} \leqslant 14990{,}6$

FIGURE 16



Avant de passer à la suite, nous allons faire un retour sur le modèle à équations multiples, équation (i), (ii) et (iii), et évoquer le problème qu'il pose pour l'induction statistique. Tout d'abord l'ajustement qu'on vient d'effectuer permet d'avoir des estimations de certains des coefficients de ces équations ; en particulier on écrira :

$$x_0 = a\,(d - 1)/cd$$
$$x_1 = a/cd$$
$$x_2 = b$$

d'où l'on peut tirer une estimation de b ($b = x_2$) et un estimateur de d ($d = (x_0 + x_1)/x_1$) ; on obtient par ailleurs un estimateur de a/c ($a/c = x_0 + x_1$), mais on ne peut fixer la valeur de chacun des coefficients a et c séparément. On ne peut surmonter cette difficulté qu'en injectant de nouvelles informations statistiques. On pense évidemment à la possibilité d'effectuer un ajustement séparé sur l'équation (i) mais dans cette équation le coefficient b est déjà estimé ($b = x_2$) ce qui pose le problème de la sur-identification du modèle.

Par contre, nous disposons d'un ajustement possible pour l'équation (iii) puisque nous avons une estimation du coefficient $d = (x_0 + x_1)/x_1$

(iii)
Calcul n° 1 :
$d = 6{,}3634$; $C_{n+1} - C_n = 6{,}3634\,(C^*_{n+1} - C_n)$
Calcul n° 2 :
$d = 5{,}8137$; $C_{n+1} - C_n = 5{,}8137\,(C^*_{n+1} - C_n)$

On a annoncé plus haut, en présentant la troisième hypothèse, que l'espérance mathématique de la quantité $(C_{n+1} - C^*_{n+1})/C_{n+1}$ constitue un estimateur naturel du pourcentage de prévenus dont la durée de prévention dépasse une année. Une estimation empirique de cette quantité peut alors s'obtenir en considérant les *moyennes* de ces variables sur les séries observées ; en particulier la moyenne correspondant à la variable C^*_{n+1} peut s'obtenir à partir des relations qu'on vient d'écrire.

	Moyenne C_n	Moyenne C^*_{+1}	Moyenne C_{n+1}	% prévenus durée préventive > 1 an
Calcul n° 1	234337,7	235649,6	242685,6	2,89 %
Calcul n° 2	233830,5	235179,1	241671,2	2,69 %

Il découle donc du modèle que, sur la période passée 1952-1970, environ 97,2 % des prévenus ont eu une durée de prévention inférieure à un an. Ce résultat, qui corrobore ce que l'on sait par ailleurs sur les durées de prévention permet de confirmer a posteriori nos hypothèses.

6. — Etude des variations saisonnières

On dispose des statistiques mensuelles des prévenus présents au 1er de chaque mois. Sur le graphique de l'évolution temporelle de cette variable on distingue très clairement des variations régulières

au cours de chaque année : en général un maximum en mai et en novembre, et un minimum très prononcé en août. On est donc amené à étudier ces cycles saisonniers autour de la tendance générale pour pouvoir moduler la prévision de l'année 1975. En effet, compte tenu de l'ampleur de ces variations, la moyenne annuelle des présents au cours de l'année, ou l'effectif des présents au 1er janvier, sont certainement des indicateurs insuffisants pour la prévision. La forme générale des évolutions nous invite à repérer ces variations trimestriellement aux dates suivantes :

1er février
1er mai
1er août
1er novembre

Le problème technique qui se pose à ce propos provient du fait que, de même que pour l'étude de la moyenne annuelle des prévenus, on ne veut pas restreindre la portée de la méthode par une hypothèse a priori sur la *tendance générale* de l'évolution. Or, les méthodes classiques d'estimation des variations saisonnières reposent pour la plupart sur l'estimation préalable (ou simultanée) d'une tendance générale — souvent supposée linéaire d'ailleurs. La méthode que l'on trouvera exposée dans l'ANNEXE IV permet, sous réserve de quelques hypothèses peu contraignantes, d'estimer les fluctuations trimestrielles sans avoir à préciser la forme de la tendance générale d'évolution ; elle consiste à appliquer l'induction statistique sur le modèle suivant :

j = 1, 2, 3, 4 indice de trimestre

P_t^j = effectif de prévenus présents au 1er jour du trimestre j de l'année t

p_t^j = part de l'effectif correspondant à la tendance générale (non déterminée) à cette date

s^j = écart saisonnier correspondant au trimestre j

r_t^j = variable aléatoire représentant les fluctuations purement aléatoires de P_t à la date indiquée

$$P_t^j = p_t^j + s^j + r_t^j \quad \begin{cases} j = 1, 2, 3, 4 \\ t = 1, 2, \ldots, n \end{cases}$$

avec $s^1 + s^2 + s^3 + s^4 = 0$

Il va de soi qu'une telle méthode, simple et rapide, ne saurait avoir les qualités et la précision d'une *analyse spectrale* de la série chronologique des prévenus présents au 1er de chaque mois — analyse qui permet de décomposer l'évolution en somme de *composantes harmoniques* dont les harmoniques à fréquence basse représenteraient justement les fluctuations saisonnières. Cependant la mise en œuvre d'une telle méthode suppose un volume assez important de calculs

complexes qui ne peuvent être présentés dans ce premier exposé des résultats. On trouvera sur le tableau qui suit les résultats des calculs effectués sur les données 1960-1971.

Estimation des fluctuations saisonnières

au 1er février	+ 183,9
au 1er mai	+ 322,8
au 1er août	— 951,4
au 1er novembre	+ 444,7

Comme on peut le constater, ces fluctuations régulières sont considérables et ne sauraient être négligées si l'on veut avoir une idée assez précise des prévisions de prévenus en 1975. D'ailleurs on peut estimer quelle part prennent les fluctuations saisonnières au cours de l'année à la variance totale des observations (voir l'annexe technique) ; en effet, la variance annuelle totale V_t et la variance V_s due aux fluctuations trimestrielles valent respectivement :

$$V_t = 22\,779\,000$$
$$V_s = 14\,528\,799$$

de sorte que 63,8 % des variations des observations sont en fait expliquées par les évolutions systématiques au cours de l'année : « creux » du mois d'août, maximum du mois de novembre, etc.

7. — **Premiers résultats de prévision des prévenus**

Le fait que l'on ait conservé provisoirement deux modèles alternatifs pour la criminalité nous amène à présenter les résultats relatifs à chacun de ces modèles ; on les trouvera ci-dessous avec les titres : calcul n° 1 et calcul n° 2. On estime tout d'abord la valeur attendue p^j_{75} au premier jour du trimestre j de l'année 1975 sur la tendance générale telle qu'elle découle des résultats numériques précédents, à laquelle on ajoute l'estimation de la fluctuation saisonnière correspondante. On calcule ensuite pour chaque début de trimestre *l'intervalle de confiance* à 95 % par application de la formule classique relative à l'estimation d'une prévision (et non de l'espérance mathématique d'une prévision — formule utilisée pour l'intervalle de confiance de $\overline{P}_{75}$). Les résultats sont rassemblés dans le TABLEAU suivant et figurés sur le graphique (FIGURE 17) qui l'accompagne. Rappelons que si l'étape de projection criminelle avait pu être achevée, elle aurait conduit à un modèle unique, et il n'apparaîtrait qu'une série de calculs pour la prévision des prévenus. On fera la même remarque à propos de la prévision des condamnés présents en prison.

CALCUL n° 1

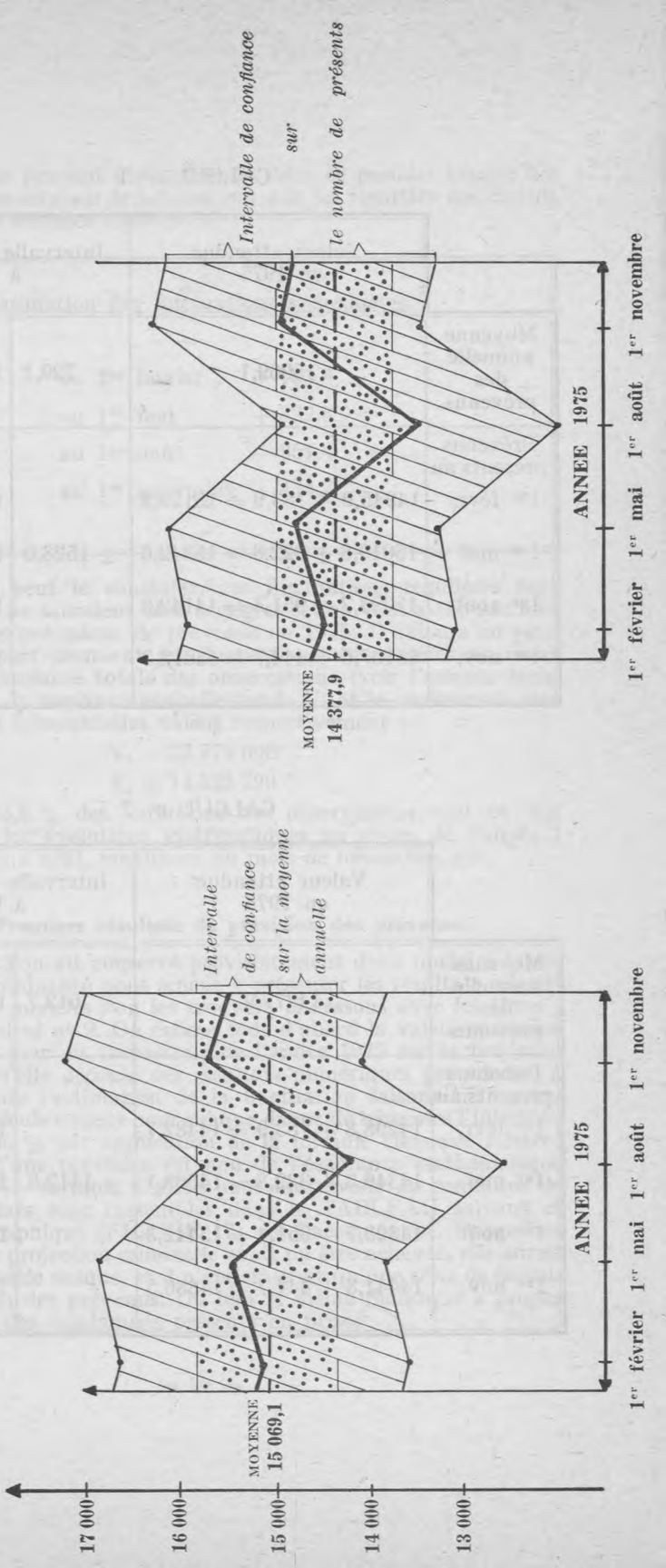
	Valeur attendue en 1975	Intervalle de confiance à 95 %	
Moyenne annuelle des prévenus	15069,1	± 729,1	14340,0/15798,2
Prévenus présents au			
1er févr.	14945,9 + 183,9 = 15129,8		13601,8/16657,8
1er mai	15019,8 + 322,8 = 15342,6	± 1528,0	13814,6/16870,6
1er août	15093,7 — 951,4 = 14142,3		12614,3/15670,3
1er nov.	15167,6 + 444,7 = 15612,3		14084,3/17140,3

CALCUL n° 2

	Valeur attendue en 1975	Intervalle de confiance à 95 %	
Moyenne annuelle des prévenus	14377,9	± 612,7	13765,2/14990,6
Prévenus présents au			
1er févr.	14298,8 + 183,9 = 14482,7		13069,9/15895,5
1er mai	14346,3 + 322,8 = 14669,1	± 1412,8	13256,3/16081,9
1er août	14393,7 — 951,4 = 13442,3		12029,5/14859,1
1er nov.	14441,8 + 444,7 = 14886,5		13473,7/16299,3

FIGURE 17

PREVISIONS DES PREVENUS POUR L'ANNEE 1975



IV. — Troisième étape

PROJECTION DES CONDAMNÉS PRÉSENTS EN PRISON

1. — Le modèle

Le modèle que nous allons utiliser, a déjà été présenté dans ces mêmes pages (Rapport Général sur l'Exercice 1969 - pp. 281-315). C'est pourquoi nous n'en rappellerons que les grandes lignes. Convenons d'appeler Q_t le nombre de condamnés présents dans les prisons au 1er janvier de l'année t. Leur présence à cette date résulte d'une condamnation à une peine d'emprisonnement ferme prononcée à une date antérieure ; plus précisément cette peine a pu être prononcée au cours de l'année $t - 1$, et donc être une part du total des condamnations C_{t-1} correspondant aux sept catégories d'infractions étudiées plus haut ; ou bien cette peine d'emprisonnement ferme a été prononcée l'année précédente, et donc être une part des C_{t-2} condamnations ; etc. Autrement dit, on peut écrire :

(1) $Q_t = a_{t-1} C_{t-1} + a_{t-2} C_{t-2} + a_{t-3} C_{t-3} + \ldots$

où les coefficients a_{t-1}, a_{t-2}, a_{t-3}, etc. représentent les parts des condamnations de l'année indiquée qui correspondent à des prisonniers présents en prison au 1er janvier de l'année t. Ceci ne constitue pas évidemment une hypothèse, mais la traduction logique du phénomène. L'hypothèse que l'on va énoncer pour définir le modèle va consister à dire que ce mécanisme de « remplissage » des prisons est, à des fluctuations aléatoires près, le même chaque année sur la période où on l'étudie. En d'autres termes la part des condamnés d'une année donnée qui se trouvent présents en prison au 1er janvier de l'année suivante est donnée par un coefficient a_1 indépendant de l'année ; de même la part des prisonniers provenant des condamnations prononcées deux ans auparavant est donnée par un coefficient a_2 indépendant de l'année considérée, etc. D'où :

PREMIÈRE HYPOTHÈSE. Soit Q_t le nombre de condamnés présents en prison au 1er janvier de l'année t, et C_t le nombre de condamnations (correspondant aux sept catégories d'infractions déjà définies) prononcées durant l'année t. Alors :

(2) $Q_t = a_1 C_{t-1} + a_2 C_{t-2} + a_3 C_{t-3} + \ldots + u_t$.
où u_t est une fluctuation aléatoire d'espérance mathématique nulle.

Cette hypothèse, qui suffirait pour caractériser complètement le modèle, ne permet pas malheureusement de réaliser une induction statistique efficace ; en effet les coefficients à estimer a_1, a_2, a_3, etc. sont trop nombreux même si l'on juge qu'à partir d'un certain délai

ils sont suffisamment faibles pour être négligés (c'est-à-dire incorporés au résidu aléatoire) ; par exemple la part des prisonniers présents en prison au 1er janvier et dont la peine a été prononcée 12 ans auparavant est certainement assez faible pour être considérée comme largement aléatoire. Pour plus de détails sur ces difficultés, on renvoie le lecteur à l'ANNEXE III et à l'article cité. L'hypothèse suivante, qui est donc une *hypothèse « technique »*, n'impose pratiquement aucune contrainte supplémentaire sur le modèle, mais permettra d'obtenir des estimations des coefficients plus précises et plus efficaces :

DEUXIÈME HYPOTHÈSE. Dans l'équation écrite (1re hypothèse) les coefficients a_1, a_2, a_3, ... prennent des valeurs choisies dans la famille à deux paramètres des DISTRIBUTIONS DE PASCAL.

Compte tenu de cette hypothèse, on verra qu'on peut transformer le modèle « à retards échelonnés » de la première hypothèse en un modèle du type « autorégressif », où l'infinité des coefficients à estimer a_1, a_2, a_3, etc. est remplacée par l'estimation de TROIS paramètres seulement. Notons à ce propos qu'il n'y a pas de relation simple entre le nombre de prévenus et le nombre de condamnés présents en prison ; d'un point de vue empirique, ceci apparaît sur la figure (18), où l'on a représenté les évolutions simultanées de ces deux variables ainsi que l'évolution de leur rapport ; on en trouve la traduction formelle en comparant le modèle des prévenus au modèle des condamnés : leurs structures diffèrent notablement.

2. — Induction statistique et résultats

Appelons a la somme des coefficients a_i :

$$a = a_1 + a_2 + a_3 + \ldots$$

Cette somme n'est pas infinie par nature (puisqu'au plus cent termes sont non nuls si on suppose qu'un individu ne vit pas plus de cent ans). Donc l'équation (2) peut s'écrire :

$$(3)\quad \begin{cases} Q_t = a\,(b_1 C_{t-1} + b_2 C_{t-2} + \ldots) + u_t \\ \text{avec } b_1 + b_2 + \ldots = 1 \end{cases}$$

D'après la deuxième hypothèse, on peut écrire :

$$b_i = (1-b)^{r+1}\left(\frac{r+i-1}{i-1}\right) b^{i-1} \qquad i = 1,\ 2,\ 3,\ \ldots$$

où b et r sont les paramètres (inconnus) des distributions de PASCAL. Dans ces conditions, on montre que l'équation (3) est équivalente à :

$$(4)\quad Q_t = \binom{r+1}{1} b\, Q_{t-1} + \binom{r+1}{2} b^2 Q_{t-2} + \ldots + (-1)^r b^{r-1} Q_{t-r-1} + a\,(1-b)^{r+1} C_{t-1} + v_t$$

où v_t est un résidu aléatoire de moyenne nulle obtenu à partir de u_t (dans l'équation (2)) par une transformation semblable. Sous la forme

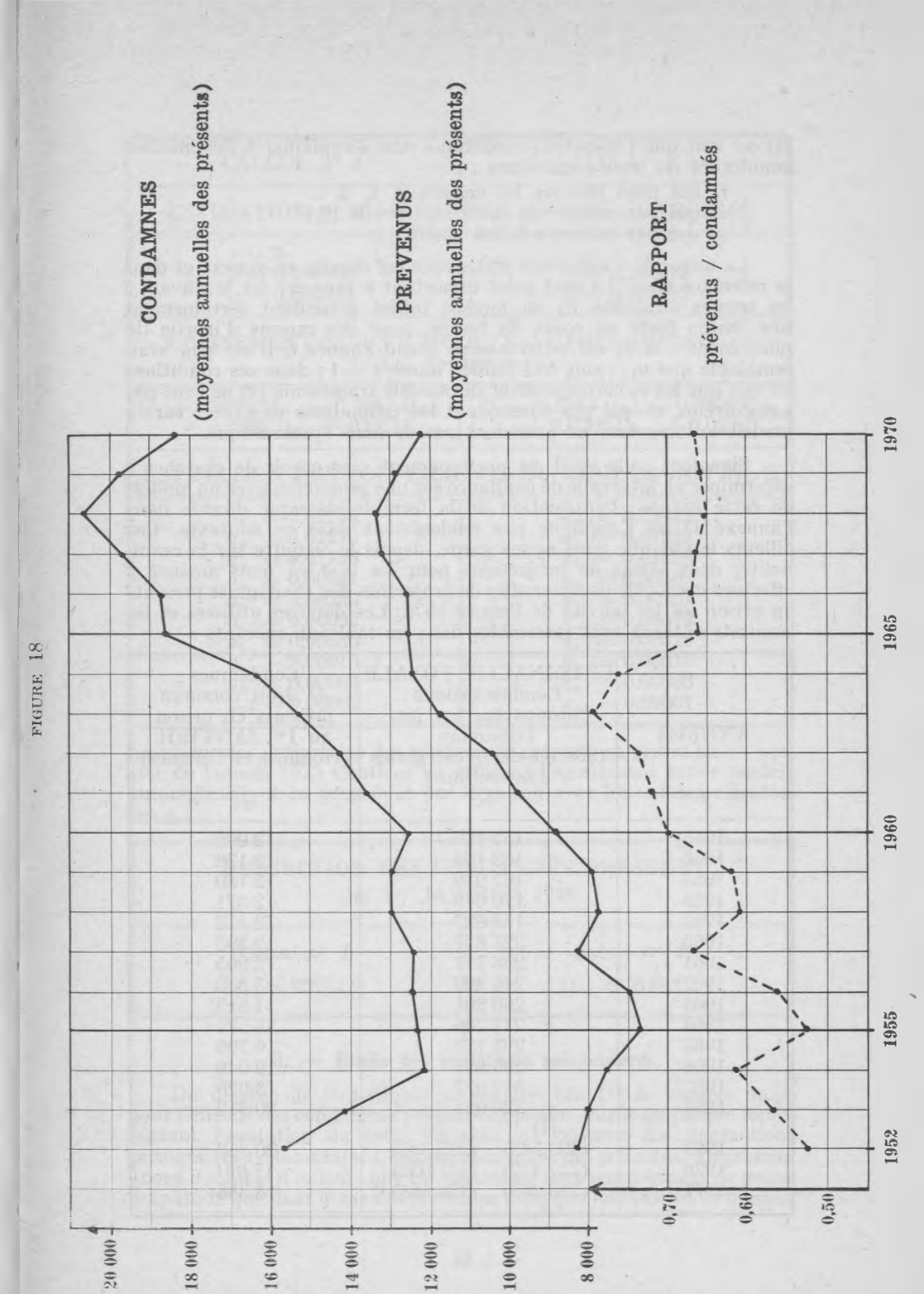


FIGURE 18

(4) on voit que l'induction statistique doit s'appliquer à l'estimation simultanée de trois paramètres :

- r, qui peut prendre les valeurs 0, 1, 2, ...
- b, qui par nature est dans l'intervalle (0,1)
- a, qui par nature est fini positif

La méthode d'induction statistique est décrite en annexe et dans la référence citée. Le seul point important à rappeler est le suivant : les termes aléatoires u_t du modèle initial présentent certainement une liaison forte au cours du temps, pour des raisons d'inertie du phénomène : si u_t est relativement grand l'année t, il est peu vraisemblable que u_{t+1} soit très faible l'année $t + 1$; dans ces conditions on sait que les v_t correspondant du modèle transformé (4) ne sont pas autocorrélés, et que par conséquent les estimations effectuées sur ce modèle autorégressif ne possèdent pas de biais systématique.

Signalons enfin qu'il est pratiquement sans espoir de chercher à déterminer un intervalle de confiance sur une projection avec un modèle de cette nature ; l'application de la formule classique donnée dans l'annexe II ne s'applique pas évidemment dans ce contexte. Par ailleurs le fait que nous ayons gardé, depuis le chapitre sur la criminalité, deux séries de projections pour les (C_t) va nous amener à effectuer deux calculs alternatifs de projection des condamnés présents en prison au 1er janvier de l'année 1975. Les données utilisées et les résultats obtenus sont rassemblés dans les tableaux suivants :

ANNÉES t	CRIMINALITÉ TOTALE Condamnations prononcées par les Tribunaux (Somme des 7 catégories d'infractions) C_t		Condamnés de droit commun présents en prison au 1er JANVIER (Hommes et Femmes) Q_t
1955	152 120		12 006
1956	152 120		12 136
1957	167 989		12 180
1958	170 649		12 371
1959	164 627		13 112
1960	223 837		12 393
1961	238 717		12 965
1962	246 800		13 830
1963	260 901		14 319
1964	277 286		15 575
1965	293 172		16 799
1966	295 897		19 049
1967	317 077		18 626
1968	320 021		20 312
	(1)	(2)	
1969	(328 749)	(325 537)	20 353
1970	(337 477)	(331 052)	17 961
1971	(346 205)	(336 568)	18 388

CALCUL n° 1	CALCUL n° 2
ESTIMATION SUR LE MODELE AUTORÉGRESSIF (4)	
$r = 2$ $b = 0{,}286708$ $a = 0{,}063312$	$r = 2$ $b = 0{,}274812$ $a = 0{,}063412$
AJUSTEMENT SUR LE MODELE AUTORÉGRESSIF (4) (Loi du processus)	
$Q_t = 0{,}8601\ Q_{t-1} - 0{,}2466\ Q_{t-2} + 0{,}0236\ Q_{t-3} + 0{,}023\ C_{t-1}$	$Q_t = 0{,}8244\ Q_{t-1} - 0{,}2266\ Q_{t-2} + 0{,}0207\ Q_{t-3} + 0{,}024\ C_{t-1}$
ESTIMATION DES COEFFICIENTS DU MODELE INITIAL (2)	
$a_1 = 0{,}02298$ $a_2 = 0{,}01976$ $a_3 = 0{,}01133$ $a_4 = 0{,}00541$ $a_5 = 0{,}00233$ $a_6 = 0{,}00093$ $a_7 = 0{,}00036$ $a_8 = 0{,}00013$ $a_9 = 0{,}00004$ $a_{10} = 0{,}00001$	$a_1 = 0{,}02418$ $a_2 = 0{,}01994$ $a_3 = 0{,}01096$ $a_4 = 0{,}00502$ $a_5 = 0{,}00207$ $a_6 = 0{,}00079$ $a_7 = 0{,}00029$ $a_8 = 0{,}00010$ $a_9 = 0{,}00003$ $a_{10} = 0{,}00001$

La valeur des projections des condamnés présents en prison au 1er janvier de l'année 1975 s'obtient en utilisant l'ajustement sur le modèle autorégressif, et en procédant par itération avec les valeurs calculées des C_t :

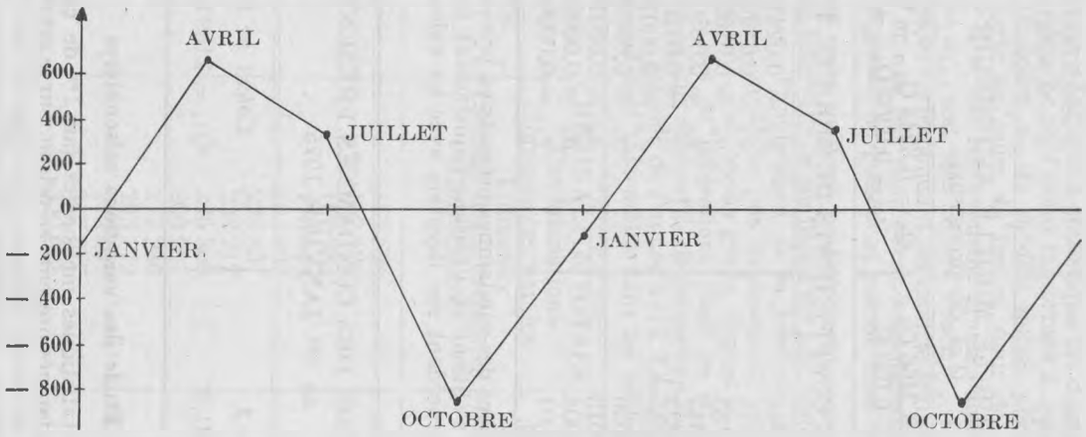
PROJECTION DES CONDAMNÉS PRÉSENTS au 1er JANVIER 1975	
Calcul n° 1	*Calcul n° 2*
$Q_{75} = 22781{,}9$	$Q_{75} = 21893{,}2$

3. — Etude des variations saisonnières

On dispose de statistiques mensuelles (au 1er de chaque mois) pour l'effectif des condamnés présents en prison. Sur le graphique représentant l'évolution de cette variable, on observe des fluctuations saisonnières systématiques, comme dans le cas des prévenus. Au premier abord il semble d'ailleurs que ces variations sont sensiblement de même amplitude que dans le cas des prévenus, mais que les maxima et minima

FIGURE 19

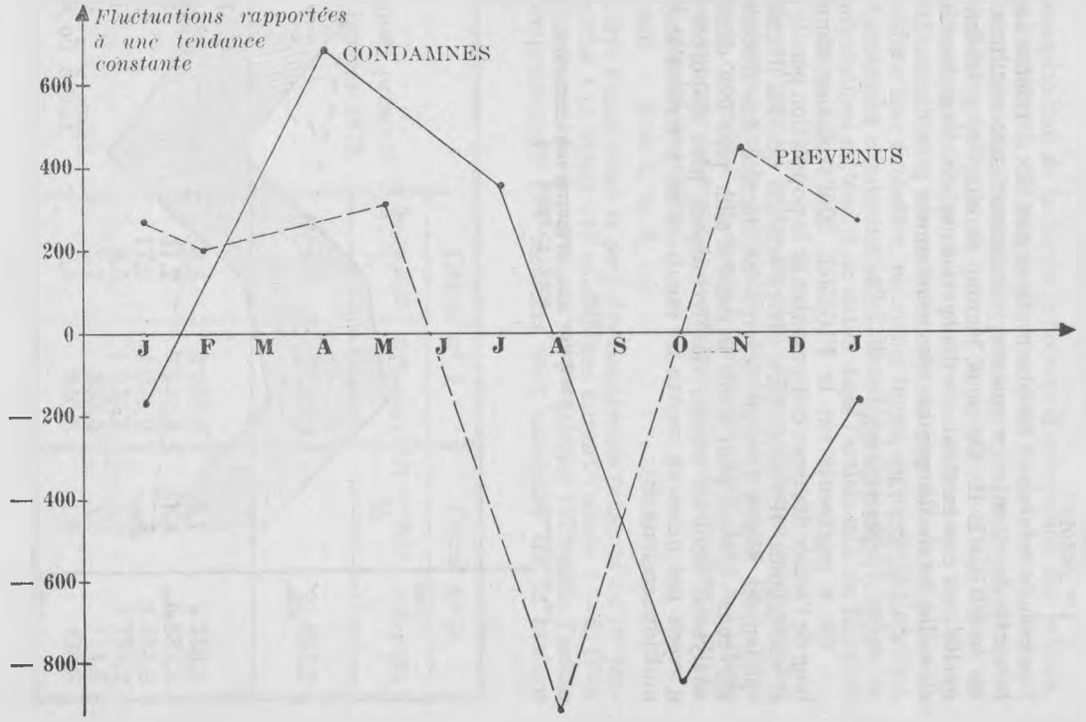
FLUCTUATIONS TRIMESTRIELLES DES CONDAMNES
(rapportées à une tendance constante)



1er janvier	— 174,7
1er avril	+ 670,6
1er juillet	+ 343,4
1er octobre	— 839,3

FIGURE 20

COMPARAISON
DES FLUCTUATIONS DE LA POPULATION PENALE
AU COURS D'UNE ANNEE



ne se produisent pas aux mêmes dates : on se propose de les étudier comme précédemment (avec la même méthode), mais aux dates suivantes qui semblent plus pertinentes :

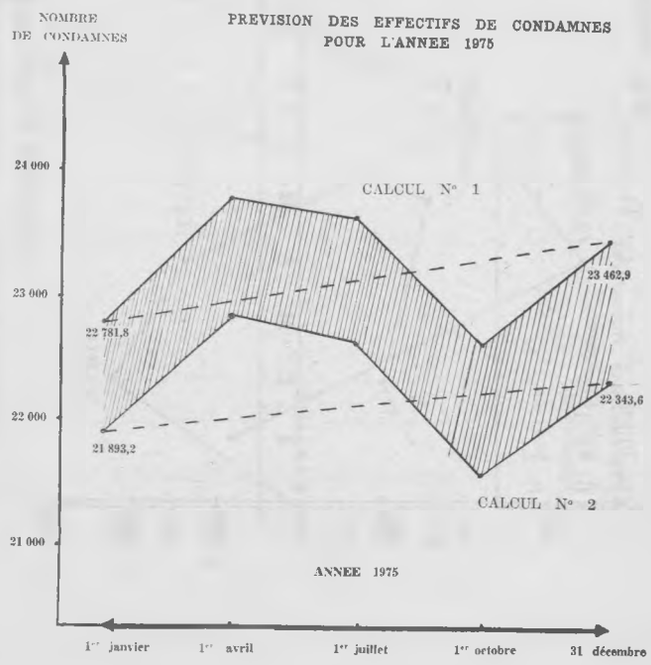
- 1er janvier
- 1er avril
- 1er juillet
- 1er octobre

Les calculs menés sur les observations des dix dernières années selon la méthode décrite en annexe, conduisent aux résultats présentés sur la FIGURE 19. On peut, comme au chapitre précédent, calculer quelle part représentent ces fluctuations saisonnières dans la variance annuelle totale du nombre des condamnés présents en prison.

$V_t = 22\,835\,130$ / $V_s = 9\,284\,313$ \ $\frac{V_s}{V_t} = 40{,}7\ \%$

On a représenté sur la FIGURE 20 l'évolution simultanée au long de l'année des deux composantes de la population pénale : prévenus et condamnés. Il semble bien sur ce schéma que les fluctuations des condamnés soient l'image légèrement décalée des fluctuations des prévenus — mais pour avoir le droit d'aller plus loin dans une telle analyse il faudrait utiliser des techniques plus élaborées que nous n'avons pas mises en œuvre ici (analyses de processus stochastiques multidimensionnels).

FIGURE 21



Finalement le modèle pour les condamnés a permis d'obtenir la prévision du nombre de présents au 1er janvier 1975, et l'étude des variations saisonnières conduit à moduler ce chiffre suivant le trimestre de l'année 1975. Les résultats sont rassemblés sur la FIGURE 21.

4. — Autres résultats

La détermination de la loi du processus d'occupation des prisons, c'est-à-dire l'induction statistique sur le modèle formel des condamnés, permet d'avoir une connaissance assez précise sur les processus d'entrées et de sorties des prisons, ainsi que sur la distribution des durées de peine à tout moment. On peut en particulier décrire sur un tableau statistique l'ensemble de ces mouvements de population et y lire de façon immédiate leur évolution, en même temps qu'on établit un pont entre les statistiques criminelles et les statistiques pénales. L'ensemble de ces manipulations est décrit en détail dans l'article cité du Rapport Général sur l'Exercice 1969.

Sans développer ici toutes les applications possibles pour enrichir notre connaissance sur la population pénale attendue en 1975, nous en donnerons cependant un exemple (correspondant dans à l'article à la « lecture sur une diagonale »). En comparant les formules (2) et (3), il apparaît que les quantités

$b_i = a_i/a \qquad (i = 1,\ 2,\ 3,\ \ldots)$

expriment très exactement la part des condamnés présents au 1er janvier de l'année t et ayant été condamnés durant l'année $t - i$. D'où la possibilité de ventiler les présents au 1er janvier 1975 selon l'année de leur condamnation. Les résultats sont consignés sur le tableau ci-dessous.

Condamnés présents le 1er janvier 1975	Calcul n° 1		Calcul n° 2	
	$b_i = a_i/a$ %	Effectifs	$b_i = a_i/a$ %	Effectifs
		22 781,9		21 893,2
Ont été jugés en :				
1974	36,3	8 269,0	38,1	8 348,2
1973	31,2	7 110,3	31,4	6 884,3
1972	17,9	4 076,9	17,3	3 784,0
1971	8,5	1 946,7	7,9	1 733,2
1970	3,7	838,4	3,3	714,7
1969 ou avant	2,4	640,6	1,9	428,8

ANNEXE I

ANALYSE DES DONNÉES

1 — Facteurs extraits d'un tableau numérique rectangulaire

Rappelons que *l'analyse des données* est une branche de la statistique mathématique distincte par nature de l'induction statistique ; pour la caractériser schématiquement on peut dire qu'il s'agit d'un ensemble de méthodes destinées à extraire et à synthétiser sous forme condensée les informations contenues dans de vastes tableaux de relevés numériques, c'est-à-dire une branche de la statistique descriptive. L'interprétation de ses résultats conduit en général à exhiber l'infrastructure d'un phénomène qui se trouve cachée et implicite dans les chiffres observés. Contrairement à l'induction statistique, l'analyse des données ne suppose *aucun modèle a priori* (voir par contre les annexes II et III). Toutes les méthodes d'analyses des données ont un tronc commun, qui est l'extraction des facteurs d'un tableau numérique rectangulaire ; elles se diversifient quant à la manière de construire ce tableau numérique à partir des observations statistiques recueillies — cette manière de construire le tableau étant dictée elle-même par la nature des observations statistiques. Nous donnerons quelques indications concernant les deux principales méthodes d'analyses utilisées dans cette étude : d'une part, l'analyse en composantes principales (PEARSON, 1901 - HOTELLING, 1933), d'autre part, l'analyse des correspondances (BENZECRI, 1964).

Considérons donc tout d'abord le problème de l'extraction des « facteurs » d'un tableau numérique Z possédant n lignes et p colonnes. On peut donner du tableau Z deux interprétations géométriques, suivant qu'on le lit en lignes ou en colonnes : il représente soit le nuage des p points-colonnes dans l'espace à n composantes R^n des lignes, soit le nuage des n points-lignes dans l'espace à p composantes R^p des colonnes. Lorsque l'on parle « d'extraire les facteurs » d'un nuage de points, on cherche en fait à ajuster de façon optimale au nuage de points un sous-espace (variété linéaire) de l'espace dans lequel il se trouve ; il y aura donc deux ajustements à effectuer suivant que l'on opère dans R^n ou dans R^p.

Considérons tout d'abord le nuage des n points-lignes dans R^p, et proposons-nous de déterminer l'axe qui passe au plus près de l'ensemble des points du nuage. Soit u le vecteur unitaire porté par cette droite :

$$u'u = 1$$

u est déterminé par la propriété de rendre maximum la somme des projections des points du nuage sur la droite, c'est-à-dire de rendre maximum la quantité :

$$u'\ Z'Z\ u$$

En d'autres termes il s'agit de trouver le vecteur unitaire u (ie. $u'u = 1$) qui maximise $u'Z'Zu$; écrivons le Lagrangien du problème :

$$u'Z'Zu - a\,(u'u - 1)$$

et annulons ses dérivées partielles par rapport aux divserses composantes de u, il vient :

$$2\,Z'Z\,u - 2\,a\,u = 0$$

c'est-à-dire

$$Z'Z\,u = a\,u$$

Autrement dit la solution u est un vecteur propre de la matrice Z'Z. En prémultipliant par u', on observe que :

$$u'\,Z'Z\,u = a\,u'u = a$$

Et puisqu'il s'agit de maximiser $u'Z'Z\,u$, on s'aperçoit que u est le vecteur propre associé à la plus grande valeur propre a de la matrice Z'Z. Notons u_1 et a_1 les quantités calculées pour déterminer cette droite et proposons-nous de trouver maintenant la droite qui avec la première définit le plan s'ajustant de façon optimale au nuage des n points de R^p. Soit donc u_2 le vecteur unitaire porté par cette droite orthogonale à u_1 ; écrivons le Lagrangien du problème :

$$u'_2\,Z'Z\,u_2 - a(u'_2\,u_2 - 1) - b\,u'_2\,u_1$$

et annulons les dérivées partielles par rapport aux composantes de u_2; il vient :

$$2\,Z'\,Z\,u_2 - 2\,a\,u_2 - b\,u_1 = 0$$

On prémultiplie par u'_1

$$2\,u'_1\,Z'\,Z\,u_2 - 2\,a\,u'_1\,u_2 - b\,u'_1\,u_1 = 0$$

Or $u'_1\,u_2 = 0$ par orthogonalité, $u'_1\,u_1 = 1$ puisque c'est un vecteur unitaire, et $u'_1\,Z'\,Z = a_1\,u'_1$ puisque c'est le vecteur propre ; donc $b\,u'_1\,u_1 = 0$, c'est-à-dire $b = 0$; par conséquent :

$$2\,Z'\,Z\,u_2 - 2\,a\,u_2 = 0$$

$$Z'\,Z\,u_2 = a\,u_2$$

Ainsi u_2 est aussi un vecteur propre associé à la seconde valeur propre a_2 de cette matrice. Le calcul se poursuit de façon semblable pour déterminer les directions u_3, u_4, etc. Remarquons que la matrice Z'Z est symétrique et généralement définie positive (exceptionnellement semi-définie positive) ; par conséquent toutes ses valeurs propres sont non-négatives et il leur correspond des vecteurs propres orthogonaux (c'est pourquoi on a cherché ci-dessus u_2 orthogonal à u_1).

Résumons ce résultat : soit Z un tableau numérique rectangulaire dont les n lignes sont considérées comme des points dans l'espace R^p des colonnes. Pour tout $q = 1, 2, \ldots p$ le sous-espace de dimensions q qui s'ajuste de façon optimale au nuage des n points est déterminé par les vecteurs propres orthogonaux $u_1, u_2, \ldots, u_q$ de la matrice Z'Z associée aux plus grandes valeurs propres $a_1, a_2, \ldots, a_q$ écrites par valeurs décroissantes.

Ce résultat se transpose immédiatement au cas où on considère le nuage des p points-colonnes dans l'espace R^n des vecteurs-lignes. Pour déterminer la droite portée par le vecteur unitaire v_1, qui s'ajuste de façon optimale au nuage des p points, il faut maximiser la somme des projections des points sur cette droite, c'est-à-dire la quantité $v_1'\,Z\,Z'\,v_1$

La suite du raisonnement est la même et on peut énoncer : soit Z le tableau numérique dont les p colonnes sont considérées comme des points dans l'espace R^n des lignes ; pour tout $q = 1, 2, \ldots, n$ le sous-espace de dimension q qui s'ajuste de façon optimale au nuage des p points est déterminé par les vecteurs propres orthogonaux $v_1, v_2, \ldots, v_q$ de la matrice ZZ' associés aux plus grandes valeurs propres $b_1, b_2, \ldots, b_q$ écrites par valeurs décroissantes.

Remarquons qu'on travaille sur Z'Z quand on raisonne dans R^p et sur Z Z' quand on raisonne dans R^n. Or c'est un résultat classique que Z Z' et Z'Z ont les mêmes valeurs propres non nulles (si on suppose, ce qui sera généralement le cas, que le tableau Z a plus de lignes que de colonnes ; i.e. $p \leqslant q$, et qu'il est de rang p, alors Z'Z et Z Z' ont toutes deux les mêmes p premières valeurs propres, la seconde matrice ayant $n - p$ valeurs propres nulles). Il y a donc certainement une relation entre les vecteurs propres $u_1, \ldots, u_p$ *et* $v_1, \ldots, v_n$ de ces deux matrices. En effet, considérons le q vecteur propre de Z'Z :

$$Z'Z\,u_q = a_q\,u_q$$

et prémultiplions par Z :

$$ZZ'\,(Z\,u_q) = a_q\,(Z\,u_q)$$

Il apparait que $Z\,u_q$ est le q^e vecteur propre de Z Z' associé à la même valeur propre a_q ; on peut donc écrire :

$$v_q = Z\,u_q \;(q = 1, 2, \ldots, p \text{ si } p < n)$$

On s'aperçoit d'autre part que si $v'_q\,v_q = 1$ alors $u'_q\,u_q = a_q$; et si $u'_q\,u_q = 1$ alors $v'_q\,v_q = 1/a_q$; pour que les deux vecteurs u_q et v_q soient simultanément *unitaires* il suffit donc de poser :

$$u_q = (1/\sqrt{a_q})Z'v_q \text{ ou } v_q = (1/\sqrt{a_q})Zu_q$$

Remarquons enfin que la connaissance des u_q, v_q associés à a_q rassemble toute l'information contenue dans Z puisque :

$$Z = \sqrt{a_1}\,v_1\,u'_1 + \sqrt{a_2}\,v_2\,u'_2 + \ldots + \sqrt{a_p}\,v_p\,u'_p$$

Enfin, on réunit le *maximum d'information* partielle contenue dans Z, et la *meilleure information* d'après la méthode même d'extraction, en ne retenant que les q premiers termes de cette somme, q étant déterminé par le degré d'approximation que l'on accepte.

En résumé on peut énoncer : soit Z un tableau numérique à n lignes et p colonnes, et supposons $p < n$. Soient $u_1, \ldots, u_q, \ldots, u_p$ les vecteurs propres unitaires de la matrice Z'Z associés aux valeurs propres $a_1, \ldots, a_q, \ldots, a_p$ écrites dans l'ordre décroissant. Pour tout $q = 1, 2, \ldots p$ on obtient les meilleures ajustements du nuage des n points-lignes dans R^p et du nuage des p points-colonnes dans R^n de la façon suivante :

On prend comme coordonnées des n points-lignes les composantes des vecteurs $Zu_1, Zu_2, \ldots, Zu_q$;

On prend comme coordonnées des p points-colonnes les composantes des vecteurs $u_1 \sqrt{a_1}$, $u_2 \sqrt{a_2}$, ..., $u_q \sqrt{a_q}$.

Alors les deux nuages de points obtenus dans R^q sont les meilleures approximations possibles du nuage des n points de R^p et du nuage des p points de R^n. Nous allons voir maintenant l'application de cette méthode de projection des nuages de points au cas de l'analyse en composantes principales et de l'analyse des correspondances.

2. — Analyse en composantes principales

On dispose au départ d'un tableau de données statistiques X à n lignes et p colonnes, où lignes et colonnes jouent des rôles dissymétriques : les lignes figurent des *individus statistiques* (par exemple les départements) alors que les colonnes figurent des *variables* mesurées sur chacun des individus statistiques. Il s'agit donc de synthétiser l'information contenue dans ce tableau avec un minimum de perte ; d'après le paragraphe précédent, on est amené à définir une notion de distance sur le tableau X pour construire les nuages de points correspondants à un tableau de même dimension Z dont on fera l'analyse comme il a été indiqué.

La distance qu'on est amené de façon naturelle à choisir doit caractériser la ressemblance ou la dissemblance entre deux individus statistiques, i.e. entre deux lignes du tableau X. Si on appelle X_i et X_j les vecteurs lignes correspondant aux lignes i et j (1) on écrira donc :

$$d(X_i, X_j) = (X_j - X_i)'(X_j - X_i)$$

Cette quantité est d'autant plus grande que les variables mesurées sur la ligne i diffèrent des mesures sur la ligne j ; elle est nulle en particulier si les variables prennent les mêmes valeurs pour les deux individus statistiques. On est donc ramené à chercher un sous-espace de projection bien choisi tel que les distances entre les projections des individus statistiques sur ce sous-espace soient des approximations correctes des distances réelles entre les n points du nuage de R^p. Par exemple, la recherche du premier axe nommé u_1 qui s'ajuste de façon optimale au nuage de points revient à résoudre le problème :

Maximiser

$$u'_1 [(X_2 - X_1)'(X_2 - X_1) + \ldots + (X^n - X_{n-1})'(X_n - X_{n-1})] u_1$$

avec $u'_1 u_1 = 1$

Or, il est facile de s'apercevoir que la quantité entre crochets n'est autre que la matrice Z'Z ou Z serait le tableau de dimension (n, p) dont le terme i,k s'écrit :

$$z_{ik} = x_{ik} - \bar{x}_k$$

où $\bar{x}_k$ est la moyenne de la variable (colonne) k. Ainsi le problème s'écrit sous la forme générale rencontrée dans le premier paragraphe

(1) Par convention i et j seront toujours des indices de lignes et k et l des indices de colonnes ; $i, j = 1, 2, \ldots, n$ et $k, l = 1, 2, \ldots, p$.

maximiser $u'\, Z'\, Z\, u$

avec $u'\, u = 1$

Le raisonnement est semblable si on cherche à représenter le nuage des p points-variables (colonnes) de l'espace R^n ; avec la même matrice Z, on s'aperçoit qu'on est amené à résoudre le problème suivant :

maximiser $v'\, Z\, Z'\, v$

avec $v'\, v = 1$

Par conséquent tous les résultats généraux s'appliquent ici. Cependant, il convient de faire une remarque importante : la quantité.

$$z_{ik} = x_{ik} - \bar{x}_k$$

dépend des unités dans lesquelles sont mesurées les variables et donc les résultats de l'analyse elle-même pourront varier selon le choix des unités de mesure, ce qui est évidemment très fâcheux. C'est pourquoi on est amené à faire l'analyse non pas sur le tableau Z ainsi défini mais sur le tableau Z dont le terme général est :

$$\boxed{z_{ik} = (x_{ik} - \bar{x}_k)/s_k}$$

où s_k est l'écart-type empirique de la variable (colonne) k. Dans ces conditions on montre aisément que la matrice Z' Z n'est autre que la MATRICE des CORRÉLATIONS des variables du tableau initial X, au facteur $1/n$ près.

Signalons encore une propriété importante qui va permettre de dégager une certaine notion de distance entre variables et individus statistiques, et par conséquent autoriser à représenter sur le même sous-espace de projection à la fois les n individus statistiques et les p variables. Calculons le coefficient de corrélation entre la $k^{\text{è}}$ variable centrée réduite (i.e. la colonne Z_k du tableau Z) et le $q^{\text{è}}$ facteur (i.e. le vecteur propre u_q associé à la valeur propre a_q de la matrice Z' Z) :

$$c(k, q) = (1/n) \sum_{i=1}^{n} \left(z_{ik} \sum_{l=1}^{p} u_{ql}\, z_{il} \right) \sqrt{\mathrm{Var}(Z_k) \cdot \mathrm{Var}(u_q)}$$

Dans cette expression $\mathrm{Var}(Z_k) = 1$ car les variables sont réduites, $\mathrm{Var}(u_q) = a_q$ puisque u_q est la q valeur propre associée à la valeur propre a_q ; enfin le terme au numérateur se simplifie pour donner $a_q\, u_{qk}$ par définition du vecteur propre ; par conséquent :

$$c(k, q) = u_{qk} \sqrt{a_q}$$

En d'autres termes, le coefficient de corrélation entre la variable Z_k et le $q^{\text{è}}$ facteur u_q est proportionnel à la k^{e} composante du facteur u_q. Par conséquent, si on porte sur le sous-espace de projection des n individus statistiques les points variables munis de ces composantes, on obtient une projection simultanée du nuage des VARIABLES et du nuage des INDIVIDUS STATISTIQUES où les proximités que l'on observe s'interprètent en termes de CORRÉLATIONS.

On part d'un tableau X d'observations de p variables sur n individus statistiques. Soit Z le tableau des variables centrées et réduites; Z' Z est, à un coefficient $1/n$ près, la matrice des corrélations entre les p variables.

On effectue l'extraction des facteurs de la matrice des corrélations (ou de la matrice Z' Z); autrement dit on cherche les vecteurs propres normés orthogonaux u_1, u_2 ... associés aux valeurs propres a_1, a_2 ... dans leur ordre décroissant.

On projette le nuage des *individus statistiques* (lignes) sur le sous-espace le mieux ajusté de dimension q, en prenant comme coordonnées du i^e individu statistique les quantités : $u'_1 X_i$; $u'_2 X_i$; ...; $u'_q X_i$.

Sur le même sous-espace on projette le nuage des p points *variables* (colonnes) en prenant comme coordonnées de la k^e variable les quantités : $u_{1k}\sqrt{a_1}$; $u_{2k}\sqrt{a_2}$; ..., $u_{qk}\sqrt{a_q}$ (en multipliant éventuellement par un facteur d'homothétie pour obtenir une dispersion de ce nuage semblable à la dispersion des individus statistiques). Alors la proximité entre individus statistiques s'interprète en terme de similitude de comportement, et la proximité entre variables et individus statistiques en terme de corrélation.

Il reste à dire un mot sur le choix de la dimension optimale q du sous-espace de projection. On peut vérifier aisément que les composantes principales ou facteurs extraits par l'analyse sont les combinaisons linéaires des variables, orthogonales entre elles, et ayant à chaque étape la variance maximum ; d'autre part, la somme des variances des composantes principales est égale à la somme des variances des variables originales. Par conséquent, on arrêtera l'extraction des facteurs au rang q si on estime que la somme des variances retenues est une part assez importante de la somme totale des variances des variables de départ. (Il n'existe pas de test systématique d'arrêt car la loi de distribution des valeurs propres n'est pas en général paramétrable de façon simple ; on opère plus commodément par simulation lorsque le résultat n'est pas évident.)

3. — **Analyse des correspondances**

Contrairement à l'analyse en composantes principales, l'analyse des correspondances s'effectue sur un tableau de données X où les lignes et les colonnes jouent un *rôle symétrique*. Plus précisément les lignes d'un côté, les colonnes de l'autre représentent deux typologies de catégories destinées à consigner des fréquences d'observations (exemple: la criminalité totale de la France est répartie en ligne par département et en colonne par catégorie d'infractions). L'objet de l'analyse se ramène à comparer les lignes entre elles (départements ayant même profil de criminalité), et à comparer les colonnes entre elles (catégories d'infractions ayant la même répartition départementale). Il s'agit donc de définir sur le tableau X des observations de départ une distance entre lignes et entre colonnes qui reflète ces notions de proximité, de sorte qu'on soit ramené à l'extraction des facteurs d'un tableau Z construit de façon adéquate.

Considérons donc le tableau des effectifs X de départ, ayant n lignes et p colonnes; i et j seront toujours des indices de ligne, k et l des indices de colonne. Convenons des notations suivantes :

$$x_{i1} + x_{i2} + \ldots + x_{ip} = x_{i.} \qquad i = 1, 2, \ldots, n$$

$$x_{1k} + x_{2k} + \ldots + x_{nk} = x_{.k} \qquad k = 1, 2, \ldots, p$$

$$x_{1.} + \ldots + x_{n.} = x_{.1} + \ldots + x_{.p} = x_{..}$$

$$p_{ik} = x_{ik}/x_{..}; \qquad p_{i.} = x_{i.}/x_{..}; \qquad p_{.k} = x_{.k}/x_{..}$$

Dans ce contexte, on prend comme mesure de la distance entre deux lignes (i) et (j) du tableau X la quantité ;

$$d(i, j) = \sum_{k=1}^{p} (1/p_{.k})\,(p_{ik}/p_{i.} - p_{jk}/p_{j.})^2$$

et de façon semblable on mesure la distance entre deux colonnes (k) et (l) du tableau X à l'aide de la formule :

$$d(k, l) = \sum_{i=1}^{n} (1/p_{i.})\,(p_{ik}/p_{.k} - p_{il}/p_{.l})^2$$

On vérifie aisément que ces formules définissent bien des distances au sens axiomatique du terme, et que d'autre part, elles permettent de mesurer effectivement la notion de proximité introduite entre lignes et entre colonnes. Nous n'insisterons pas sur les propriétés intéressantes de ces distances sinon pour évoquer les conséquences fondamentales de la propriété « *d'équivalence distributionnelle* ». On remarquera, en effet, que si par exemple deux lignes (i) et (j) sont identiques, i.e. si les deux points de R^p sont confondus, on ne change pas le reste des résultats en les remplaçant par un point unique affecté de la somme des poids des deux points. Cette propriété a pour effet de rendre les résultats de l'analyse pratiquement INDÉPENDANTS des choix des catégories en ligne ou en colonne. Ainsi, par exemple, une analyse départementale donnera les mêmes résultats qu'une analyse régionale, ou une analyse communale, ou même une analyse où les unités géographiques seraient regroupées selon un critère autre que géographique. Par conséquent, les résultats de l'analyse seront dans une très large mesure *intrinsèques*, et c'est sans doute là ce qui fait la puissance de la méthode.

Il reste à montrer comment ces notions de distance permettront de construire le tableau Z sur lequel se fera l'extraction des facteurs. On montre facilement que la maximisation des distances projetées des n points du nuage de R^p conduit au problème :

maximiser $u'\,Z'\,Z\,u$

avec $u'u = 1$

et où Z est le tableau à n lignes et p colonnes de temre général :

$$z_{ik} = (p_{ik} - p_{i.}\,p_{.k}) \,/\, \sqrt{p_{i.}\,p_{.k}}$$

Quant au problème de la projection du nuage des p points colonnes de R^n, il se ramène à la maximisation de la quantité $v'\,Z\,Z'\,v$ avec $v'\,v = 1$. Il apparaît cependant ici une simplification notable si l'on remarque que les nuages de points sont contenus dans le simplexe

orthogonal au vecteur de composantes $\sqrt{p_{.k}}$ (pour $k = 1, ..., p$). Or ce vecteur est le vecteur propre de Z' Z correspondant à la valeur propre 1 ; l'élimination de ce vecteur parasite dans l'analyse conduit à travailler sur le tableau Z de terme général.

$$z_{ik} = p_{ik} / \sqrt{p_{i.}\, p_{.k}}$$

Les matrices Z' Z et Z Z' ont alors $p - 1$ valeurs propres non nulles correspondant aux facteurs extraits du tableau.

RÉSUMÉ DE LA MÉTHODE D'ANALYSE DES CORRESPONDANCES

On part d'un tableau X ayant n lignes et p colonnes, tableau de contingence indiquant la répartition d'un effectif donné d'observations dans diverses catégories selon une typologie écrite en ligne et une autre typologie écrite en colonne.

On construit le tableau Z de terme général $Z_{ik} = p_{ik} / \sqrt{p_{i.}\, p_{.k}}$ et on extrait les facteurs de la matrice Z' Z ; *i.e.* on calcule les vecteurs propres u_1, u_2, ..., u_q (avec $q \geq p\text{-}1$) correspondant aux valeurs propres non nulles écrites dans l'ordre décroissant.

On projette le nuage des n *catégories en ligne* sur le sous-espace le mieux ajusté de dimension q, en prenant comme coordonnées du i^e point-ligne les quantités :

$$\sum_{k=1}^{p} \left(p_{ik} / p_{i.} \sqrt{p_{.k}} \right) u_{1k} \, ; \, ... \, ; \sum_{k=1}^{p} \left(p_{ik} / p_{i.} \sqrt{p_{.k}} \right) u_{qk}$$

sur le même sous-espace on projette le nuage des p *catégories en colonne* en prenant comme coordonnées de la $k^è$ colonne les quantités :

$$\sqrt{a_1 / p_{.k}}\, u_{1k} \, ; \, ... \, ; \sqrt{a_q / p_{.k}}\, u_{qk}$$

Alors les notions de proximité entre profils de répartition en ligne, entre profils de répartition en colonne et entre catégories-lignes et catégories-colonnes sont représentées par les proximités entre les points projetés dans le sous-espace à q dimensions.

On peut faire les mêmes remarques que dans le cas de l'analyse en composantes principales pour ce qui est du choix de la dimension q du sous-espace de projection. Signalons que dans la plupart des applications il est suffisant de projeter dans l'espace des deux premiers facteurs extraits, mais qu'exceptionnellement un seul facteur suffit dans certains cas (les tests sont effectués ici aussi par simulation de matrices Z aléatoires pour déterminer la probabilité qu'une valeur propre soit significativement supérieure à sa valeur dans un tableau purement aléatoire).

ANNEXE II

INDUCTION STATISTIQUE SUR LE MODÈLE LINÉAIRE

1. — Les hypothèses du modèle ; problème d'estimation

Notre propos est de rappeler ici quelques résultats classiques sur l'induction statistique propre au modèle linéaire (résultats qui ont été abondamment utilisés au cours de la recherche) et de signaler à l'occasion quelques erreurs d'interprétation qu'il est malheureusement courant de rencontrer dans les travaux hâtifs. Ces précisions justifieront les multiples précautions de langage que le lecteur aura rencontrées dans cette étude.

On considère un vecteur aléatoire Y à n composantes, et on suppose que pour tout $i = 1, 2, ..., n$ la variable Y_i obéit à la loi suivante :

$$Y_i = X_{i1}\, b_1 + X_{i2}\, b_2 + ... + X_{ip}\, b_p + u_i$$

où les X_{ij} sont des constantes connues, alors que les b_j sont des paramètres inconnus (mais non aléatoires) ; enfin u_i est une variable aléatoire non observable. On note X la matrice de dimension (n, p) des X_{ij}, b la matrice de dimension $(p, 1)$ des b_j, et u la matrice aléatoire de dimension $(n, 1)$ des u_i. Avec ces notations les n relations sur les Y_i s'écrivent sous une forme matricielle simple qui constitue la première hypothèse du modèle :

HYPOTHÈSE 1 : $Y = X\, b + u$

On notera E l'opérateur « espérance mathématique » ; la matrice des variances-covariances du vecteur Y est par définition :

$$E\,[(Y - X\, b)\,(Y - X\, b)'] = E\,(u\, u')$$

Les résultats de l'induction statistique sur le modèle linéaire feront intervenir suivant le cas une ou plusieurs des hypothèses qui suivent :

HYPOTHÈSE 2 : les b_j ne sont soumis à aucune contrainte.

HYPOTHÈSE 3 : la matrice X n'est pas aléatoire.

HYPOTHÈSE 4 : le rang de X est égal à p (avec $p < n$).

HYPOTHÈSE 5 : $E\,(u) = 0$ c'est-à-dire $E\,(Y) = X\, b$.

HYPOTHÈSE 6 : en appelant I_n la matrice unité d'ordre n et s^2 un scalaire positif, on a $E\,(u\, u') = s^2\, I_n$.

HYPOTHÈSE 7 : s^2 ne dépend pas de b.

HYPOTHÈSE 8 : le vecteur u suit une loi normale à n dimensions notée $N\,(0, s^2\, I_n)$.

Nous ne commenterons pas ces hypothèses, bien qu'il faille être parfaitement maître de leur contenu et de leur signification avant de songer à utiliser un modèle linéaire (1). Le premier problème posé est d'obtenir par induction statistique, au vu d'UNE observation du vecteur Y, des estimateurs du vecteur b et du scalaire s^2 qui jouissent de bonnes propriétés. Pour effectuer ces estimations, il s'avère que la méthode dite des moindres carrés possède les meilleures justifications comme on va le rappeler rapidement (le lecteur suppléera aux commentaires omis).

Soit $\hat{b}$ un estimateur de b ; le vecteur $e = Y - X\hat{b}$ est appelé le résidu ; l'estimateur $\hat{b}$ donné par la méthode des moindres carrés est celui qui minimise la somme des carrés des résidus, c'est-à-dire la quantité $e'e$. On démontre les propriétés suivantes (en appelant $\hat{e} = Y - X\hat{b}$) :

— Sous les hypothèses (1), (2), et (4), on a $\hat{b} = (X'X)^{-1}X'Y$

— Sous les hypothèses (1), (3), (4) et (5) on a $E(\hat{b}) = b$

— Sous les hypothèses (1), (3), (4), (5) et (6) on a

$E[(\hat{b} - b)(\hat{b} - b)'] = s^2(X'X)^{-1}$

$E(\hat{e}) = 0$

$E(\hat{e}(\hat{b} - b)') = 0$

$E(\hat{e}'\hat{e}) = (n - p)s^2$

Une conséquence importante de ces propriétés est que la quantité

$$\hat{s}^2 = \frac{1}{n-p}\hat{e}'\hat{e}$$

est un estimateur sans biais de s^2, autrement dit $E(\hat{s}^2) = s^2$, et par la suite la matrice $\hat{S} = \hat{s}^2(X'X)^{-1}$ est un estimateur sans biais de la matrice de variances-covariances de $\hat{b}$. Citons enfin les propriétés qui assurent la véritable justification de ces estimateurs, et dont la première est le célèbre théorème de GAUSS-MARKOV :

— Sous les hypothèses (1) à (6) l'estimateur des moindres carrés $\hat{b}$ minimise le carré de toute somme pondérée des erreurs dans la classe des estimateurs linéaires sans biais. De plus c'est l'unique estimateur linéaire sans biais qui ait cette propriété.

— Sous les mêmes hypothèses on démontre en théorie de la décision statistique que $\hat{b}$ est ADMISSIBLE et MINIMAX dans la classe des estimateurs linéaires.

— Sous les hypothèses (1) à (8), $\hat{b}$ est aussi l'estimateur du MAXIMUM de VRAISEMBLANCE ; de plus le couple $(\hat{b}, \hat{s}^2)$ constitue un RÉSUMÉ EXHAUSTIF de Y.

---

(1) Dans un rapport plus détaillé, nous expliciterons ce qu'impliquent en termes de phénomène criminel ces hypothèses appliquées au modèle de la criminalité utilisé dans cette étude.

— Sous les hypothèses (1) à (8), $\hat{b}$ suit une loi normale $N(\hat{b}, \hat{s}^2(X'X)^{-1})$; $(n - p)\frac{\hat{s}^2}{s^2}$ suit une loi de KHI. 2 à $n - p$ degrés de liberté ; enfin $\hat{b}$ et $\hat{s}^2$ sont indépendants.

Nous ne ferons que quelques remarques sur ces résultats fondamentaux. Tout d'abord le coefficient de corrélation multiple qui est défini par :

$$R^2 = \frac{(X\hat{b} - \bar{Y})'(X\hat{b} - \bar{Y})}{(Y - \bar{Y})'(Y - \bar{Y})}$$

où $\bar{Y}$ est le vecteur de dimensions $(n, 1)$ dont chaque composante est la moyenne des observations Y_i, n'est intervenu à aucune étape du développement de la méthode d'estimation ; de fait ce coefficient n'a pas de signification intéressante contrairement à ce que l'on croit souvent ; nous aurons l'occasion de revenir sur cette remarque importante à propos de l'étude de la validité de l'estimation. Une autre erreur commune consiste à tirer des conclusions hâtives du fait que $\hat{b}$ est un estimateur du maximum de vraisemblance, et de croire que cet estimateur sera de ce fait convergent lorsque le nombre des observations Y_i augmentera ; il n'en est évidemment rien puisqu'on ne dispose que d'UNE observation (du vecteur Y) et non d'un échantillon de taille n d'une même variable aléatoire. Excepté sous des hypothèses supplémentaires et fort restrictives, la multiplication des composants Y_i n'améliore pas la qualité de l'estimation (seuls sont affectés le seuil de confiance et la puissance des tests qu'on effectue sur le modèle). Rappelons que ce n'est pas une considération de cette nature qui nous a incités à travailler sur des données départementales (d'ailleurs nombreuses) plutôt que sur des séries chronologiques.

2. — Tests d'hypothèses linéaires, précision de l'ajustement

Qu'il s'agisse d'étudier la signification individuelle d'un coefficient de régression estimé $\hat{b}_j$, ou de tester un sous-ensemble de coefficients, on se convaincra aisément qu'il s'agit toujours de tester si q relations linéaires indépendantes entre les b_j sont vraies ou non. Autrement dit ces problèmes admettent la formulation générale suivante :

Soit Q une matrice connue de dimension (q, p) et de rang q avec $q \leqslant p$; soit r un vecteur colonne connu de dimension $(q,1)$; il s'agit de tester l'hypothèse :

(H_o) $Qb = r$

contre l'hypothèse alternative :

(H_1) $Qb \neq r$

On démontre que, sous les hypothèses (1) à (8) et dans le cas où l'hypothèse nulle (H_o) est vraie, alors la variable aléatoire :

$$F = \frac{n-p}{q}\frac{(Q\hat{b} - r)'(Q(X'X)^{-1}Q')^{-1}(Q\hat{b} - r)}{\hat{e}'\hat{e}}$$

suit une loi de FISHER à q et $n - p$ degrés de libertés. Par conséquent, le test s'effectue de la façon suivante : on cherche dans la table de FISHER correspondante la valeur F (a) qui a une probabilité a d'être dépassée (on choisit en général $a = 0{,}05$; si la valeur F calculée dans l'expression développée ci-dessus est inférieure à F (a) alors on ne peut pas rejeter l'hypothèse nulle (H_o). On démontre que ce test jouit de bonnes propriétés ; en particulier il est uniformément le plus puissant (UMP) de seuil a, quel que soit a, parmi tous les tests qui laissent le problème invariant ; de plus il maximise la puissance minimale du test. C'est donc un test de ce type qu'on effectue pour savoir si un ou plusieurs coefficients de régression sont significativement nuls (c'est-à-dire pour savoir si les variables exogènes correspondantes affectent ou non la valeur de la variable endogène).

Il faut faire ici une remarque très importante. Supposons qu'on veuille tester l'hypothèse (H_o) que TOUS les coefficients de régression sont nuls, contre l'hypothèse alternative (H_1) affirmant qu'il n'en est rien. Il est facile de s'assurer que le test repose alors sur la quantité F développée plus haut qui se simplifie pour donner

$$F = \frac{n - p - 1}{p} \cdot \frac{R^2}{1 - R^2}$$

F ne dépend que du coefficient de régression multiple R. En d'autres termes un test sur R ne peut renseigner que sur un point : est-ce que TOUS les coefficients de régression sont significativement nuls, ou bien en existe-t-il UN au moins qui soit significativement différent de zéro (1)? Enfin, et pour être parfaitement explicite, la valeur du coefficient de régression multiple et les tests sur ce coefficient ne renseignent EN AUCUN CAS sur le plus ou moins bon AJUSTEMENT de la régression. Le seul moyen de juger de la précision de l'ajustement est donc d'étudier les variances (ou les écarts-types) des coefficients de régression ; chercher à maximiser le coefficient de régression multiple n'est justifié par aucun argument théorique (pas plus que la démarche qui consiste à sélectionner de préférence comme variables exogènes celles qui sont corrélées fortement avec la variable endogène : comme on l'a rappelé dans le chapitre des méthodes, les variables exogènes doivent être choisies A PRIORI, c'est-à-dire non pas n'importe comment mais après réflexions sur les observations, réflexions appuyées en général sur des analyses de données).

Pour terminer ce paragraphe signalons enfin une difficulté qui est source également d'erreurs d'interprétation. Il arrive souvent que l'HYPOTHÈSE (4) du modèle soit vérifiée, mais soit mal vérifiée, autrement dit il existe une relation linéaire approximative entre les variables exogènes : on dit qu'il y a MULTICOLLINÉARITÉ. Alors

(1) Il s'avère d'ailleurs qu'il est très difficile d'obtenir des réponses négatives à ce test au seuil habituel 0,05 ; autrement dit, quelques soient les variables exogènes aussi mal choisies soient-elles, il existe presque toujours un coefficient de régression qui n'est pas significativement nul. La valeur du coefficient de corrélation n'apporte donc aucune information intéressante.

le déterminant de X' X est presque nul et par conséquent les éléments de $(X' X)^{-1}$ sont très grands, et donc aussi ceux de la matrice des variances-covariances de l'estimateur $\hat{b}$: les coefficients de régression semblent très imprécis. En particulier, il est probable qu'on trouvera que les coefficients de régression sont individuellement nuls de façon significative, alors que paradoxalement on ne pourra pas rejeter l'hypothèse qu'ils sont tous nuls simultanément. Le paradoxe n'est qu'apparent, et traduit le fait qu'au vu de l'observation du vecteur Y on ne peut pas séparer l'influence des variables qui sont collinéaires. Cependant les conditions du théorème du GAUSS-MARKOV sont vérifiées et la méthode des moindres carrés jouit encore ici des propriétés optimales qu'on a évoquées. Par conséquent, et contrairement à une coutume répandue, on n'a pas le droit et rien ne justifie d'éliminer les variables collinéaires sous prétexte d'obtenir des estimations plus précises. Le modèle a priori étant choisi et justifié une fois pour toutes, une telle procédure ne peut conduire qu'à des erreurs de spécification (voir le chapitre introductif sur les méthodes, paragraphe 4).

3. — Les procédures de régression pas à pas

Considérons maintenant le problème suivant qui est très important dans la pratique. Supposons que des analyses de données nous aient conduits à écrire un modèle a priori de la réalité qui contienne n variables exogènes ; en général n est grand et on peut soupçonner certaines variables exogènes de ne pas influencer significativement la variable endogène (remarquer que ceci ne remet pas en cause la validité du modèle a priori choisi, l'abandon de certaines variables faisant alors partie de l'induction statistique sur ce modèle). Il serait donc fort utile de disposer d'une méthode systématique pour éliminer ou introduire progressivement les variables exogènes dans le cadre permanent du modèle a priori, au vu de leur contribution à l'explication de la variable endogène. La difficulté fondamentale réside dans le fait que si une variable est significative dans un sous-ensemble fixé de variables exogènes, elle peut fort bien ne plus l'être si on ajoute ou retranche une ou plusieurs variables dans ce sous-ensemble. A l'heure actuelle aucune méthode scientifiquement fondée ne permet de résoudre ce problème, sinon celle qui est exclue évidemment et qui consisterait à effectuer les 2^n régressions possibles.

Dans notre recherche, nous utilisons accessoirement un algorithme de régression pas à pas, décrit par M.A. EFROYMSOM (voir « Méthodes Mathématiques pour Calculateurs Arithmétiques », DUNOD - 1965). Le critère de sélection des variables exogènes repose sur l'évaluation de la contribution de la variable à la variance totale de la régression intermédiaire. Il est facile de s'assurer que le test repose en fait sur le coefficient de régression multiple et consiste à rechercher à chaque pas la maximisation de ce coefficient. En effet, la quantité F développée plus haut est égale à la quantité :

$$F = \frac{n - p}{q} \cdot \frac{V_q - \hat{V}}{\hat{V}}$$

où $\hat{V}$ est la variance totale des observations, et V_q la variance minimale dans la régression ne contenant que q des p variables exogènes. Une telle procédure est donc en toute rigueur incorrecte ; elle a cependant l'avantage d'être facilement programmable et de fournir malgré tout quelques indications sur l'importance de certaines variables exogènes ; on prendra soin pourtant de ne pas accorder trop d'importance aux résultats intermédiaires.

Devant le manque d'arguments théoriques, la prudence requierrait que l'on essaie de confirmer les résultats de la méthode utilisée en les confrontant à une ou plusieurs autres méthodes ; la lourdeur de la tâche de programmation des calculs nous en a empêchés. On aurait pu en effet utiliser des algorithmes alternatifs de régression pas à pas reposant par exemple sur les critères suivants :

— choisir la variable exogène de telle sorte qu'on augmente en valeur absolue le coefficient de corrélation partielle du plus grand nombre de variables ;

— choisir la variable de telle sorte que la somme des carrés des coefficients de corrélation *partielle* soit la plus grande possible à chaque étape, etc.

Remarquons que le critère retenu (maximation du coefficient de corrélation *multiple*) est équivalent à rendre l'écart-type du terme aléatoire u_i le plus grand possible.

4. — Prévisions avec le modèle linéaire

On considère toujours le modèle $Y = X\,b + u$ qui satisfait aux HYPOTHÈSES (1) à (6), et on suppose connues de nouvelles valeurs de chaque variable exogène, c'est-à-dire un vecteur-ligne à p composantes.

$X'_0 = (X_{01}, X_{02}, ..., X_{0p})$

Pour que l'induction statistique puisse s'appliquer il est nécessaire de faire de nouvelles hypothèses concernant le comportement du phénomène. L'hypothèse la plus naturelle est la suivante :

HYPOTHÈSE (9) : La prévision Y_0 suit la loi $Y_0 = X'_0\,b + u_0$ où u_0 est une variable aléatoire de moyenne nulle, de variance s^2, et non correlée avec le vecteur aléatoire u.

Cette hypothèse ne fait que traduire la permanence de la structure du phénomène. Le problème de la prévision est soit d'estimer une prévision de l'espérance mathématique de Y_0, soit plus communément d'estimer Y_0 lui-même et ceci au vu d'UNE observation du vecteur Y ; la prévision sera donc une variable aléatoire notée Z_0, fonction de Y, de X et de X'_0. Nous allons rappeler rapidement pourquoi l'induction statistique conduit à prendre comme prévision la quantité :

$Z_0 = X'_0\,\hat{b}$ avec $\hat{b} = (X'\,X)^{-1}\,X'\,Y$

Restreignons-nous au cas habituel où on veut prévoir la valeur de Y_0 et non l'espérance mathématique de cette valeur. Il est naturel

de chercher l'estimateur de Z_0 de cette variable aléatoire qui minimise le risque d'erreur de prévision. On démontre que, sous les hypothèses (1) à (6) et (9), la quantité

$Z_0 = X'_0\,\hat{b}$

où $\hat{b}$ est l'estimateur des moindres carrés de b, est l'unique prévision linéaire et centrée de Y_0 qui minimise le risque d'erreur (ie. la quantité Var $(Z_0 - Y_0)$ dans la classe des prévisions linéaires et centrées de Y_0, pour toute valeur des variables exogènes X'_0. L'erreur de prévision, non observable, est alors $(X'_0\,\hat{b} - Y_0)$ et sa variance est égale à :

$s^2\,[1 + X'_0\,(X'\,X)^{-1}\,X_0]$

Dans la pratique le problème de la prévision de Y_0 se présente sous une forme légèrement différente dans la mesure où on cherche à déterminer un INTERVALLE de CONFIANCE pour la prévision de Y_0. Il est alors nécessaire de supposer vérifiée une HYPOTHÈSE plus forte que l'HYPOTHÈSE (9) :

HYPOTHÈSE (10)

La prévision Y_0 suit la loi $Y_0 = X'_0\,b + u_0$ où u_0 est une variable NORMALE N $(0, s^2)$ non correlée avec le vecteur aléatoire u.

Alors sous les hypothèses (1) à (8) et (10), il apparaît que la variable aléatoire $(X'_0\,\hat{b} - Y_0)$ suit une loi normale et on montre que la quantité

$$t = \frac{X'_0\,\hat{b} - Y_0}{(\hat{s}^2 + X'_0\,\hat{S}\,X_0)^{1/2}} \quad (\text{avec } \hat{S} = \hat{s}^2\,(X'\,X)^{-1})$$

suit une loi de STUDENT à $(n - p)$ degrés de liberté. Cette propriété permet de construire un intervalle de confiance pour la prévision de Y_0 de la façon suivante : ayant choisi un seuil de confiance a (en général $a = 0{,}95$) on lit dans la table de STUDENT la valeur $t\,(a)$ telle que Prob $(|\,t\,| \leqslant t\,(a)) = a$. *L'intervalle de confiance de seuil a pour* Y_0 est alors donné par

$X'_0\,\hat{b} - t\,(a)\,(\hat{s}^2 + X'_0\,\hat{S}\,X_0)^{1/2} \leqslant Y_0 \leqslant X'_0\,\hat{b} + t\,(a)\,(\hat{s}^2 + X'_0\,\hat{S}\,X_0)^{1/2}$

Notons que l'utilisation du modèle linéaire « géographique » pour simuler la régression autorise difficilement à admettre une hypothèse de normalité des résidus (hypothèses 8 et 10), de sorte qu'il serait illusoire pour nous de calculer des intervalles de confiance des prévisions criminelles.

Terminons enfin en signalant le piège que l'on rencontre à propos des prévisions lorsqu'il y a (comme c'est le cas dans notre étude) des phénomènes de collinéarité des variables exogènes. On a signalé que cette collinéarité entraînait une certaine imprécision sur les coefficients de régression, imprécision qu'on ne pouvait éviter sans risquer des erreurs de spécification graves sur le modèle. Qu'en est-il des prévisions ? Supposons pour fixer les idées qu'il y ait collinéarité entre les deux premières variables exogènes X_1 et X_2 :

$X_1 = c\,X_2$ (approximativement)

S'il se trouve que cette relation est également vérifiée pour les valeurs X_{01} et X_{02} des variables exogènes, c'est-à-dire si $X_{01} = c\,X_{02}$ (appro-

ximativement) alors le calcul montre qu'il est pratiquement indifférent d'effectuer la prévision avec le modèle complet mais imprécis ou avec le modèle tronqué de la variable collinéaire. Cependant si, comme il faut s'y attendre en général, la liaison n'est plus vérifiée pour les valeurs X_{01} et X_{02} des variables exogènes, alors il est facile de s'apercevoir que l'utilisation du modèle tronqué, mais plus précis, conduit à une prévision systématiquement biaisée et donc fausse. Il est alors nécessaire de travailler sur le modèle initial bien qu'il soit imprécis ; la prévision sera alors également assez imprécise en général, ce qui traduira avec raison le fait que les observations ne permettent pas de savoir ce qui arrive lorsque X_{01} diffère de $c\,X_{02}$.

ANNEXE III

INDUCTION STATISTIQUE SUR LE MODÈLE A RETARDS ÉCHELONNÉS

1. — Le modèle et les hypothèses

L'induction statistique sur les modèles à retards échelonnés n'est pas, comme c'est pratiquement le cas pour les modèles linéaires, un problème parfaitement résolu, et on mettra l'accent dans cette note sur diverses difficultés qu'il est encore difficile de surmonter. Le lecteur pourra se reporter également à une présentation plus sommaire du point de vue théorique, faite dans le Rapport Général sur l'Exercice 1969 de l'Administration Pénitentiaire (« Recherche sur les processus d'entrée et d'occupation des prisons », pages 281-315).

On dispose de deux séries chronologiques d'observations des valeurs de deux variables P_t et C_t où P_t est la variable endogène du phénomène. Le modèle de simulation a priori stipule que P_t obéit à la loi suivante :

$$P_t = a_1 C_{t-1} + a_2 C_{t-2} + \ldots + u_t$$

où u_t est une variable aléatoire non observable ; par ailleurs, le nombre des coefficients a_1, a_2, a_3, etc. entrant dans la définition du modèle n'est généralement pas connu à l'avance. (On ignore à partir de quand la variable exogène n'a plus d'influence sur la variable endogène.) Il est clair que l'induction statistique sera très pauvre si on ne spécifie pas davantage le modèle. Et la difficulté essentielle réside dans le choix raisonné de ces hypothèses supplémentaires.

Ces hypothèses porteront tout d'abord, comme dans le cas du modèle linéaire, sur la loi de distribution des perturbations aléatoires u_t : sont-elles corrélées, obéissent-elles à un processus temporel, sont-elles normales? Il importe d'avoir des idées précises sur les erreurs u_t car la nature du modèle est telle que ces termes aléatoires n'ont pas en général de spécification simple comme dans le modèle linéaire. On verra qu'on est amené cependant à effectuer des simplifications pour lesquelles la théorie ne sait pas dire à l'heure actuelle si elles affectent beaucoup l'exactitude des résultats. D'autre part, et ceci est spécifique à ce type de modèle, on sera amené à faire des hypothèses a priori sur la suite des coefficients a_1, a_2, a_3, ..., soit pour en limiter le nombre, soit pour leur imposer des conditions dictées par la connaissance que l'on a du phénomène. Peut-être pourra-t-on affirmer que la suite doit être décroissante, ou encore qu'elle a une forme analytique donnée, etc.

2. — Problèmes de l'induction statistique directe

Supposons qu'on ait de bonnes raisons de croire que l'influence de la variable exogène sur la variable endogène ne se fait plus sentir au bout de p unités de temps. Cette hypothèse sur les coefficients a_i conduit à écrire que la variable aléatoire P_t obéit dans ce cas à la loi :

$$P_t = a_1 C_{t-1} + a_2 C_{t-2} + \ldots + a_p C_{t-p} + u_t$$

Une telle hypothèse évite d'avoir à formuler des restrictions supplémentaires sur les coefficients a_i, car on reconnaît en fait ici un modèle linéaire classique tel qu'on l'a utilisé dans l'annexe II. En particulier, on est amené de façon naturelle à employer la méthode d'estimation des moindres carrés. Il se présente cependant quelques particularités.

Si les perturbations aléatoires sont supposées satisfaire aux hypothèses classiques énoncées dans l'annexe II, alors la méthode des moindres carrés jouit de toutes les propriétés que l'on a vues et, en particulier, conduira à des estimations sans biais des coefficients inconnus a_1, a_2, ..., a_p. En fait, il y a tout lieu de craindre avec de telles séries chronologiques que la valeur de l'aléa u_t au temps t soit liée à la valeur prise au temps $t - 1$, et peut-être aussi au temps $t - 2$, etc. On dispose heureusement de procédures de tests pour étudier si ces aléas sont indépendants dans le cadre du modèle linéaire en particulier le test simple de DURBIN et WATSON (1950). En cas de dépendance, on sait que la méthode des moindres carrés conduit à des estimateurs biaisés des coefficients a_i. On applique dans ce cas certaines procédures d'estimation dans les modèles à erreurs liées lorsque l'on sait spécifier a priori cette liaison (mais nous n'en parlerons pas ici, n'ayant pas eu à les utiliser dans les premiers calculs présentés ici).

La liaison des erreurs n'est donc pas sans doute une difficulté insurmontable. Une autre particularité réside dans un phénomène que nous avons déjà évoqué à propos du modèle linéaire : c'est la multicollinéarité ; elle apparaît presque inéluctablement ici du fait que la série chronologique C_t présente très généralement une certaine régularité. On a signalé que ce phénomène entraînait une définition imprécise des coefficients de régression a_1, a_2, ..., a_p (qui possèdent en effet de grands écarts-types). Alors que dans le cas du modèle linéaire usuel cette difficulté est apparue insurmontable, il est possible ici d'y remédier en partie dans la mesure où la nature du problème et la connaissance du phénomène permettent de prendre en compte des informations (des hypothèses) supplémentaires sur ces coefficients. Mais il s'avère qu'en pratique cette procédure conduit à des algorithmes fort complexes qui font hésiter à les employer. Nous avons pu cependant utiliser une telle méthode d'induction directe sur un modèle à « retards » échelonnés sur deux années liant l'effectif de population en détention préventive aux effectifs de condamnés des deux années suivantes. La même méthode n'est évidemment plus applicable dans le cas plus général des effectifs présents en prison car alors de véritables retards s'échelonnent sur un plus grand nombre d'unités de temps, et ce nombre est INCONNU (et il n'existe pas de méthode théorique correcte pour l'estimer).

C'est pourquoi il a été nécessaire de développer une autre procédure d'induction.

3. — Hypothèse sur les coefficients et transformation du modèle

On se trouve maintenant dans le cas où rien ne permet de limiter a priori le nombre des coefficients a_i ; le modèle s'écrit :

$$P_t = a_1 C_{t-1} + a_2 C_{t-2} + \ldots + u_t$$

On démontre aisément que tout modèle à retards échelonnés de ce type peut s'écrire sous une forme AUTOREGRESSIVE, c'est-à-dire ne dépendant que d'une variable exogène (par exemple C_{t-1}) et faisant intervenir des formes retardées de la variable endogène (par exemple P_{t-1}, ou P_{t-1} et P_{t-2}, etc.).

Cependant si on ne prend pas de précautions supplémentaires, la forme autorégressive du modèle peut être aussi complexe que sa formulation initiale.

Remarquons que les coefficients a_i du modèle vont décroître et tendre vers 0 aussi vite que décroissent et tendent vers 0 les proportions des peines en fonction de leur durée ; on est donc assuré en pratique que la série des a_i est convergente ; appelons a sa somme :

$$a = a_1 + a_2 + a_3 + \ldots$$

et soient

b_1, b_2, b_3, ... les coefficients définis par :

$b_1 = a_1/a$; $b_2 = a_2/a$; etc.

Alors le modèle s'écrit de façon équivalente.

$$P_t = a (b_1 C_{t-1} + b_2 C_{t-2} + \ldots) + u_t$$

avec $b_1 + b_2 + \ldots = 1$

Les b_i (comme les a_i) sont certainement tous positifs et leur somme est alors égale à 1, de sorte qu'ils se présentent comme *une distribution de probabilité* sur les entiers 1, 2, 3, ... Cette dernière remarque nous conduit de façon naturelle à chercher à remplacer la suite (éventuellement infinie) des coefficients du modèle par une famille paramétrée de distributions de probabilité, de telle sorte que la détermination des paramètres (peu nombreux en général) permette de calculer ex-post tous les coefficients inconnus. Il s'avère que la famille des distributions de probabilité dites de PASCAL présente, lorsqu'on fait varier des paramètres, une très grande variété de formes possibles ; l'induction statistique devra permettre d'estimer ses paramètres, et par conséquent d'estimer par cet artifice toute la suite des coefficients du modèle. Nous allons voir au passage que les calculs sont particulièrement simples si on travaille sur la forme transformée autorégressive du modèle. Rappelons auparavant les HYPOTHESES du modèle a priori que nous avons rencontrées au cours de cette discussion :

HYPOTHÈSE (1) :

$P_t = a_1 C_{t-1} + a_2 C_{t-2} + \ldots + u_t$ ou u_t est une variable aléatoire non observable.

HYPOTHÈSE (2) :

La suite des a_i a tous ses termes non négatifs, et converge vers a, de sorte que le modèle s'écrit aussi :

$P_t = a\,(b_1\,C_{t-1} + b_2\,C_{t-2} + \ldots) + u_t$ où la somme des b_i vaut 1.

HYPOTHÈSE (3) :

Les valeurs des b_i sont choisies parmi les distributions de PASCAL à deux paramètres b et r :

$$b_i = (1-b)^{r+1}\binom{r+i-1}{i-1}\,b^{i-1}$$

avec $i = 1, 2, 3, \ldots$; $r = 0, 1,2, \ldots$; $0 \leqslant b \leqslant 1$

Avec ces hypothèses on peut transformer le modèle initial pour l'écrire sous forme autorégressive (les calculs intermédiaires, simples mais fastidieux, sont omis) ; il vient :

$$P_t = \binom{r+1}{1}\,b\,P_{t-1} - \binom{r+1}{2}\,b^2\,P_{t-2} + \ldots + (-b)^r\,P_{t-r+1} + a\,(1-b)^{r+1}\,C_{t-1} + v_t$$

Cette expression dépend de trois paramètres : les quantités a, b et r ; de plus elle contient un terme aléatoire v_t qui est le transformé de u_t, et s'en déduit donc par une transformation analogue qui n'est pas développée ici. Quoiqu'il en soit il apparaît que les hypothèses du modèle a priori entraînent que l'effectif des présents en prison à la date t dépend de façon linéaire des effectifs présents aux dates $(t-1)$, $(t-2)$, ..., $(t-r+1)$, auxquels s'ajoute en pourcentage donné de la population condamnée durant l'année $(t-1)$.

Ce résultat qui exprime la loi du processus d'occupation des prisons, prend des formes particulièrement simples pour les premières valeurs du paramètre r :

CAS $r = 0$

Dans ce cas $b_i = (1-b)\,b^{i-1}$, et le modèle s'écrit :

$P_t = b\,P_{t-1} + a\,(1-b)\,C_{t-1} + v_t$

avec $0 \leqslant b \leqslant 1$; les coefficients a_i du modèle initial sont déterminés par : $a_i = a\,(1-b)\,b^{i-1}$.

CAS $r = 1$

Dans ce cas $b_i = i\,(1-b)^2\,b^{i-1}$, et le modèle s'écrit :

$P_t = 2\,b\,P_{t-1} - b^2\,P_{t-1} + a\,(1-b)^2\,C_{t-1} + v_t$

avec $0 \leqslant b \leqslant 1$; les coefficients a_i du modèle initial sont alors déterminés par $a_i = i\,a\,(1-b)^2\,b^{i-1}$.

CAS $r = 2$

Dans ce cas $b_i = i\,(1+i)/2.(1-b)^3\,b^{i-1}$, et le modèle devient :

$P_t = 3\,b\,P_{t-1} - 3\,b^2\,P_{t-2} + b^3\,P_{t-3} + a\,(1-b)^3\,C_{t-1} + v_t$

avec $0 \leqslant b \leqslant 1$; les coefficients a_i du modèle initial sont alors déterminés par $a_i = i\,(1+i)/2.a\,(1-b)^3\,b^{i-1}$

4. — Induction statistique sur le modèle transformé

L'HYPOTHÈSE (3) du modèle a priori et la transformation sous forme autorégressive ont donc permis de remplacer l'estimation de la suite des coefficients a_i du modèle initial par l'estimation des *trois paramètres a, b* et *r* du modèle résultant. On a de cette manière surmonté la difficulté née de l'ignorance du nombre de coefficients a_i non nuls ; mais le modèle obtenu possède encore des particularités embarrassantes.

Il s'avère tout d'abord que l'induction statistique *ne permet pas* d'estimer *simultanément* les trois paramètres inconnus a, b et r ; par contre, on sait estimer par une méthode rigoureusement fondée les deux paramètres a et b lorsque la valeur de r est connue. Cette remarque nous conduit à opérer empiriquement de la façon suivante : on estimera les paramètres a et b pour diverses valeurs de r ; pour chaque valeur de r, on calculera les valeurs de P_t obtenues en remplaçant dans le modèle les coefficients inconnus par leurs estimations ; enfin, on déterminera la série des P_t calculés la plus proche (au sens de la distance du KHI-2) de la série des observations ; la valeur correspondante de r sera prise comme estimation de ce paramètre, de sorte que le modèle sera alors complètement déterminé. (Pour les détails sur la méthode d'estimation voir l'article cité sur le processus d'occupation des prisons.)

Reste le point le plus délicat du modèle, et qui conditionne les propriétés des estimateurs trouvés : à quelle loi obéissent les résidus aléatoires u_t du modèle initial, et quelle est la transformée de cette loi pour les résidus v_t du modèle autorégressif? Autrement dit quelle hypothèse a priori est-il raisonnable de faire (compte tenu de notre connaissance des phénomènes) sur la loi du terme aléatoire u_t? La difficulté naît du résultat classique suivant : même si les variables aléatoires u_t du modèle à retards échelonnés constituent un processus purement aléatoire (et a fortiori si elles sont autocorrelées) alors les variables aléatoires transformées v_t du modèle autorégressif sont, elles, nécessairement autocorrélées et par conséquent les estimations du modèle possèdent un biais asymptotique qui n'a en principe aucune raison d'être faible. Par exemple, si on suppose que les u_t obéissent à un processus purement aléatoire, alors les v_t obéissent à un processus de moyenne mobile dont le corrélogramme dépend du paramètre a. Cependant on peut démontrer que *l'auto corrélation des* v_t *est d'autant plus faible que l'autocorrélation des* u_t *est plus grande* ; or, on a de bonnes raisons de supposer que les u_t sont fortement autocorrélés dans le modèle à retards échelonnés, ne serait-ce qu'à cause de l'inertie observée dans les évolutions du phénomène. Par conséquent, et bien qu'elle ne puisse être testée rigoureusement, il semble qu'on soit autorisé à ajouter au modèle a priori l'hypothèse suivante :

HYPOTHÈSE (4)

L'autocorrélation des termes aléatoires u_t du modèle à retards échelonnés est telle que les termes aléatoires v_t du modèle transformé autorégressif suivent un processus purement aléatoire.

Dans ces conditions l'induction statistique appliquée au modèle satisfaisant aux HYPOTHÈSES (1) à (4) conduit à des estimations non biaisées et asymptotiquement convergentes lorsque le nombre d'observations augmente. Sans être optimales, ces estimations présentent suffisamment de qualités pour être acceptées. Dès lors, le modèle peut être utilisé à des fins de PROJECTION selon la méthodologie rappelée dans le premier chapitre de cette note : l'hypothèse de constance de la structure du phénomène sur la période considérée conduit à une estimation de la valeur attendue possédant formellement les mêmes qualités que les estimateurs du modèle.

ANNEXE IV

ESTIMATION DES FLUCTUATIONS TRIMESTRIELLES

1. — Le modèle

On dispose d'une série chronologique d'observations $P^j{}_t$ faites tous les trimestres ($j = 1, 2, 3, 4$) pendant n années ($t = 1, 2, ..., n$). La représentation graphique permet de supposer qu'il existe, autour de la tendance générale, des fluctuations trimestrielles régulières qu'on se propose d'estimer. Il existe de nombreuses méthodes classiques pour effectuer cette estimation, mais elles supposent en général l'estimation préliminaire d'une tendance générale, qui d'ailleurs est souvent choisie linéaire. La méthode que nous présentons ci-dessous est moins restrictive dans la mesure où elle permet une estimation DIRECTE des composantes trimestrielles sans spécification préalable de la tendance à long terme. C'est une méthode d'application rapide et simple mais il est clair qu'elle ne saurait avoir la puissance d'une ANALYSE SPECTRALE (estimation des fonctions de répartition et de densité spectrales, analyse de la fonction de covariance, représentation spectrale de la série, spécification des harmoniques, etc.), l'analyse spectrale de séries utilisées dans cette étude sera menée dans une étape ultérieure de la recherche. Nous énonçons ci-dessous les hypothèses du modèle :

HYPOTHÈSE 1 : L'observation $P^j{}_t$ contient une composante trimestrielle s^j qui s'ajoute à la tendance générale $p^j{}_t$, et à un résidu aléatoire non observable $r^j{}_t$:

$$\boxed{P^j{}_t = p^j{}_t + s^j + r^j{}_t} \quad \begin{matrix} j = 1, 2, 3, 4, \\ t = 1, 2, ..., \end{matrix}$$

Par définition de la composante saisonnière, on a :

$$\boxed{\sum_{j=1}^{4} s^j = 0}$$

HYPOTHÈSE 2 : La moyenne des résidus aléatoires est nulle à tout moment de l'année :

$$\boxed{\sum_{t=1}^{n} r^j{}_t = 0} \quad \text{pour } j = 1, 2, 3, 4,$$

De plus la moyenne ANNUELLE des résidus est « orthogonale » au déroulement du temps :

$$\sum_{t=1}^{n} \left(t \left(\sum_{j=1}^{4} r^j{}_t \right) \right) = 0$$

HYPOTHÈSE 3 : (Hypothèse de comportement à long terme). Les moyennes des observations trimestrielles de la tendance générale forment une PROGRESSION ARITHMETIQUE ; de plus la moyenne des écarts trimestriels successifs pour la tendance générale est proportionnelle à la pente de la droite d'ajustement linéaire.

Pour donner une présentation formelle de l'hypothèse 3 appelons k_t la quantité :

$$k_t = 2t - (n+1) \qquad t = 1, 2, \ldots, n$$

La droite d'ajustement linéaire sur la moyenne annuelle de la tendance générale aurait pour pente :

$$[3/n\,(n^2-1)] \sum_{t=1}^{n} \left\{ k_t \left(\sum_{j=1}^{4} p^j{}_t \right) \right\}$$

D'où la relation :

$$\sum_{t=1}^{n} \left(p^j{}_{t+1} - p^j{}_t \right) = 3/n\,(n_2 - 1).(n/8). \sum_{t=1}^{4} \left\{ k_t \left(\sum_{j=1}^{4} p^j{}_t \right) \right\}$$

Soit encore :

$$\sum_{t=1}^{n} (p^j{}_{t+1} - p^j{}_t) = (3/8\,(n^2-1). \sum_{t=1}^{n} \left\{ k_t \left(\sum_{j=1}^{4} p^j{}_t \right) \right\}$$

Nous allons montrer que ces hypothèses a priori, assez peu restrictives, suffisent pour effectuer l'estimation des fluctuations trimestrielles s^j sans spécification de la tendance générale autre que ce que contient l'hypothèse 3.

2. — Induction statistique

Consignons les observations $P^j{}_t$ dans un tableau à 4 colonnes et à n lignes : chaque ligne est une année, et chaque colonne un trimestre. Appelons S^j les sommes en colonnes (par trimestre), et P_t les sommes en ligne (par année) ; soit T le total général des observations :

$$T = \sum_{t=1}^{n} \sum_{j=1}^{4} P^j{}_t = \sum_{t=1}^{n} P_t = \sum_{j=1}^{n} S^j$$

D'après les hypothèses 1 et 2, on a :

$$S^j = \sum_{t=1}^{n} P^j{}_t + n\,s^j$$

et

$$P_t = \sum_{j=1}^{4} P^j{}_t + \sum_{j=1}^{4} r^j{}_t$$

En tenant compte de l'hypothèse 3, on a :

$$\sum_{t=1}^{n} k_t P_t = \sum_{t=1}^{n} \left\{ k_t \left(\sum_{j=1}^{4} p^j{}_t \right) \right\}$$

Ces trois dernières relations permettent d'écrire :

$$n\,(s^j - s^{j+1}) = (S^j - S^{j+1}) + (3/8\,(n^2-1)]. \sum_{t=1}^{n} k_t P_t$$

D'où l'on tire la valeur de l'estimation de s^j :

$$s^j = (1/4\,n).(4\,S^j - T) + [3\,(5 - 2j)/16\,n\,(n^2-1)]. \sum_{t=1}^{n} k_t P_t$$

3. — Remarque

L'ensemble des observations $P^j{}_t$ présente une certaine variance annuelle totale V_T qui d'après, le modèle doit être partagée entre une certaine part due aux fluctuations trimestrielles V_S, et une autre part due à la tendance générale et aux fluctuations autour de cette tendance. On vérifie aisément que l'on a :

— Variance totale ANNUELLE :

$$V_T = \sum_{t=1}^{n} \sum_{j=1}^{4} (P^j{}_t)^2 - (1/4). \sum_{t=1}^{n} (P_t)^2$$

— Variance due aux fluctuations trimestrielles dans l'année :

$$V_S = 2 \sum_{j=1}^{4} S^j s^j - n \sum_{j=1}^{4} (s^j)^2$$

Par conséquent $100.(V_S/V_T)$ représente le pourcentage de la variance annuelle des observations expliquée par les fluctuations trimestrielles.

ANNEXE V

CATÉGORIES D'INFRACTIONS

1. — **Délinquance violente et banale contre les biens**

— *Crimes*

Vol qualifié
Recel qualifié
Autres crimes
Incendie volontaire
Autres destructions

— *Délits*

Vol qualifié et vol
Recel qualifié et recel
Grivèlerie et Filouteries
Dégradation de monuments et destructions diverses
Incendie volontaire

— *Contraventions de 5e Classe*

Destruction d'arbres appartenant à autrui
Destruction d'animaux
Inondation des chemins ou propriété d'autrui

2. — **Délinquance astucieuse contre les biens**

— *Crimes*

Détournements de deniers publics
Fausse monnaie
Faux en écritures publiques
Faux en écritures privées
Abus de confiance qualifié
Banqueroute frauduleuse
Extorsion de signature

— *Délits*

Escroquerie
Abus de confiance qualifié
Abus de confiance
Abus blanc seing
Détournement
Faux en écritures publiques ou privées
Faux et usage de faux
Fausse monnaie
Banqueroute frauduleuse ou simple

Fraudes commerciales
Contrefaçon
Action illicite sur marché
Prix illicites
Liberté des enchères
Publicité mensongère
Faux certificat qualité
Récompense industrielle
Appellation d'origine
Défaut carte professionnelle
Démarcharge
Autres infractions économiques
Loyers
Usure
Valeurs mobilières
Autres infractions commerciales
Infractions banque et bourse
Change
Autres infractions fiscales
Rétention de précomptes
Jeux et paris
Loteries

— *Contraventions de 5e Classe*

Rétention de précompte
Défaut de carte professionnelle

3. — Atteintes volontaires contre les personnes

— *Crimes*

Meurtre - Assassinat
Parricide
Empoisonnement
Coups mortels et autres blessures volontaires
Infanticide
Autres crimes

— *Délits*

Meurtre, empoisonnement
Blessures volontaires
Coups à enfant
Violences et voies de fait

— *Contraventions*

Violences et voies de fait (5e classe)

4. — Atteintes involontaires contre les personnes

— *Délits*

Homicide involontaire
Blessures involontaires (circulation et autres)

— *Contraventions de 5e Classe*

Blessures involontaires

5. — Infractions contre les mœurs

— *Crimes*

Viol, attentat à la pudeur sur mineur
Viol, « « sur adulte

— *Délits*

Attentat à la pudeur sur adulte, (mineur)
Outrage public à la pudeur
Homosexualité
Proxénétisme et aide à la prostitution
Pornographie
Outrage aux bonnes mœurs

— *Contraventions de 5e Classe*

Accès mineurs dans certains établissements

6. — Infractions aux règles de la circulation

— *Délits*

Conduite en état d'ivresse
Course sans autorisation
Entrave à la circulation
Véhicules et équipement
Condition de circulation des véhicules
Conduite sans permis
Défaut d'assurances
Coordination des transports

— *Contraventions de 5e Classe*

Coordination des transports

7. — Infractions contre la chose publique

— *Crimes*

Violences à fonctionnaire
Détournements de deniers publics
Association malfaiteurs
Autres crimes contre la chose publique
Sûreté de l'état

— *Délits et contraventions de 5e Classe*

Administration de substances vénéneuses
Médecine et professions paramédicales
Autres professions réglementées
Entrave à la liberté du travail
Ivresse

Débits de boissons - Alcoolisme
Infractions à la loi sur les inhumations
Contraventions à l'article 216 code S.P.
Contraventions à l'article 279 code S.P.
Exercice illégal de la profession de sage-femme
Exercice illégal de la profession d'infirmier
Exercice illégal de la profession de masseur
Réglementation des substances vénéneuses
Contraventions au code du travail
Dénonciation calomnieuse
Secret professionnel
Menaces
Diffamation, injures
Refus de porter secours
Violation de domicile, bris de clôture
Rebellion, violences, outrage à fonctionnaire
Faux témoignage et subornation
Non dénonciation
Aide à malfaiteur
Recel de malfaiteur
Correspondance de détenu
Evasion
Interdiction de séjour
Refus d'un service dû
Vagabondage, mendicité
Nomade
Expulsion, séjour des étrangers
Armes et explosifs
Police des chemins de fer
Sûreté de l'État
Attroupements, réunions, manifestation
Associations
Elections
Autres délits de presse
Atteinte au crédit de la nation
Postes
Délits fluviaux
Délits maritimes
Outrage à service public
Port illégal de décorations
Infractions commises par un officier d'état civi
Infractions reatives aux actes de naissance
Contraventions forestières
Police des chemins de fer
Défaut de carte de séjour

ANNEXE VI (A)

DEPARTEMENTS	POPULATION MASCULINE DE 18 ANS ET PLUS		CRIMINALITE BRUTE		TAUX DE CRIMINALITE = Criminalité / Population × 1 000	
	1962	1968	1962	1968	1962	1968
01 Ain	112 362	118 973	1 092	1 625	9,7186	13,6586
02 Aisne	163 028	170 856	2 911	3 996	17,8558	23,3881
03 Allier	131 836	136 209	1 635	1 994	12,4018	44,6393
04 Alpes-de-Provence	34 053	37 998	532	697	15,6227	18,3431
05 Hautes-Alpes	31 034	33 518	511	680	16,4658	20,2876
06 Alpes-Maritimes	223 588	262 830	4 899	6 520	21,9108	24,8069
07 Ardèche	84 895	89 930	608	1 026	7,1618	11,4089
08 Ardennes	96 561	99 214	2 310	1 822	23,9227	18,3643
09 Ariège	49 792	50 766	213	587	4,2778	11,5629
10 Aube	84 379	91 786	1 466	1 806	17,3740	19,6762
11 Aude	95 467	99 050	1 118	1 498	11,7109	15,1237
12 Aveyron	103 420	99 765	679	1 089	6,5655	10,9157
13 Bouches-du-Rhône	433 725	513 381	7 711	11 574	17,7785	22,5447
14 Calvados	147 131	163 399	3 655	4 337	24,8418	26,5424
15 Cantal	58 940	59 429	668	622	11,3336	10,4663
16 Charente	110 360	112 261	1 208	1 538	10,9460	13,7002
17 Charente-Maritime	155 853	163 343	2 792	3 204	17,9143	19,6152
18 Cher	101 880	107 463	1 407	1 614	13,8104	15,0191
19 Corrèze	84 448	85 967	695	1 273	8,2299	14,8080
20 Corse	65 620	78 168	1 326	2 619	20,2073	33,5048
21 Côte-d'Or	128 219	142 252	2 720	2 447	21,2137	17,2019
22 Côtes-du-Nord	163 707	168 263	1 711	3 141	10,4516	18,6672
23 Creuse	59 362	57 513	463	641	7,7996	11,1453
24 Dordogne	130 351	133 460	1 691	1 814	12,9727	13,5921
25 Doubs	124 658	140 285	2 103	2 342	16,8702	16,6946
26 Drôme	103 334	116 371	1 918	3 112	18,5612	26,7421
27 Eure	115 383	125 845	1 820	2 499	15,7736	19,8578
28 Eure-et-Loir	91 578	101 899	1 453	1 850	15,8663	18,1552
29 Finistère	246 403	256 835	2 430	3 813	9,8619	14,8461
30 Gard	148 013	164 884	2 040	2 627	13,7826	15,9324
31 Haute-Garonne	205 572	243 102	1 905	3 643	9,2668	14,9855
32 Gers	65 281	65 816	747	805	11,4428	12,2311
33 Gironde	311 012	341 924	5 727	4 689	18,4141	13,7136
34 Hérault	175 482	205 315	1 909	3 098	10,8786	15,0890
35 Ille-et-Vilaine	194 167	210 436	3 378	4 327	17,3974	20,5621
36 Indre	86 360	86 568	995	1 597	11,5215	18,4479
37 Indre-et-Loire	129 827	146 522	1 897	2 480	14,6118	16,9258
38 Isère	248 744	260 251	3 629	4 484	14,5893	17,2295
39 Jura	74 962	79 270	912	1 278	12,1662	16,1221
40 Landes	90 193	96 774	826	1 328	9,1581	13,7227
41 Loir-et-Cher	83 269	91 322	1 034	1 461	12,4176	15,9983
42 Loire	231 675	245 991	2 803	4 666	12,0988	18,9682
43 Haute-Loire	71 301	71 174	824	786	11,5566	11,0434
44 Loire-Atlantique	249 298	273 674	4 192	5 204	16,8152	19,0153
45 Loiret	130 308	148 172	2 164	3 386	16,6068	22,8518
46 Lot	52 158	52 959	639	679	12,2512	12,8212
47 Lot-et-Garonne	94 882	101 944	1 128	1 744	11,8885	17,1074
48 Lozère	28 899	27 647	442	316	15,2946	11,4298
49 Maine-et-Loire	171 005	183 170	2 374	3 029	13,8826	16,5366
50 Manche	137 242	142 817	1 735	2 896	12,6419	20,2777
51 Marne	143 238	163 362	3 131	4 409	21,8587	26,9891
52 Haute-Marne	66 922	71 204	1 410	1 438	21,0693	20,1955
53 Mayenne	78 379	80 720	1 282	1 485	16,3564	18,3969
54 Meurthe-et-Moselle	222 229	235 167	4 263	4 299	19,1829	18,2806
55 Meuse	69 120	69 407	1 374	1 517	19,8785	21,8566
56 Morbihan	170 346	176 484	1 989	3 217	11,6762	18,2283
57 Moselle	311 403	317 358	7 025	7 564	22,5592	23,8343
58 Nièvre	83 847	85 726	1 339	1 936	15,9696	22,5836
59 Nord	732 586	782 831	13 635	17 076	18,6121	21,8131
60 Oise	157 101	178 625	3 258	3 960	20,7383	22,1693
61 Orne	87 994	92 447	1 373	2 145	15,6033	23,2025
62 Pas-de-Calais	425 390	443 323	6 560	7 689	15,4211	17,3440
63 Puy-de-Dôme	175 335	194 234	2 219	2 859	12,6630	14,7194
64 Basses-Pyrénées	157 203	172 898	2 031	2 983	12,9196	17,2529
65 Hautes-Pyrénées	72 912	78 955	760	1 585	10,4235	20,0747
66 Pyrénées-Orientales	89 062	99 706	1 401	2 316	15,7306	23,2283
67 Bas-Rhin	256 295	275 284	4 077	4 356	16,9074	15,8237
68 Haut-Rhin	183 676	194 302	4 077	4 753	22,1967	24,4619
69 Rhône	379 379	449 751	5 063	7 649	13,3455	17,0072
70 Haute-Saône et Territoire-de-Belfort	103 870	110 948	1 643	2 319	15,8178	20,9017
71 Saône-et-Loire	181 241	189 150	2 574	3 050	14,2021	16,1248
72 Sarthe	136 756	147 238	2 763	2 894	20,2039	18,2969
73 Savoie	90 563	100 095	1 532	2 398	16,9164	23,9572
74 Haute-Savoie	112 670	128 976	2 122	2 307	18,8338	17,8870
75 Seine et Seine-et-Oise	2 771 076	3 058 439	47 307	58 273	17,0717	19,0532
76 Seine-Maritime	327 908	361 302	7 425	7 688	22,6435	21,2786
77 Seine-et-Marne	175 779	206 955	3 215	6 028	18,2900	29,1271
79 Deux-Sèvres	106 783	109 297	792	1 487	7,4169	13,6051
80 Somme	192 108	167 926	2 829	4 128	14,7261	24,5823
81 Tarn	110 412	116 258	1 096	1 453	9,9265	12,4981
82 Tarn-et-Garonne	59 680	63 642	645	997	10,8076	15,6658
83 Var	168 378	197 595	3 148	5 321	18,6960	26,9288
84 Vaucluse	105 462	122 846	1 944	3 301	18,4332	26,8710
85 Vendée	128 594	135 547	1 012	1 531	7,8697	11,2950
86 Vienne	110 817	115 293	1 325	2 058	11,9568	17,8502
87 Haute-Vienne	116 663	122 442	746	1 254	6,3945	10,2416
88 Vosges	118 326	125 226	1 517	1 982	12,8205	15,8274
89 Yonne	92 084	98 879	1 841	2 213	19,9926	22,3809
FRANCE ENTIERE	15 580 560	16 896 599	246 788	330 021	15,839	18,939

CATEGORIES SOCIO-PROFESSIONNELLES MASCULINES EN 1962

Pourcentages par rapport à la population masculine
de 15 ans et plus, active et non active

ANNEXE VI (B)

DEPARTEMENTS	POPULATION MASCULINE de 15 ans et plus	0 (*)	1 (*)	2 (*)	3 (*)	4 (*)	5 (*)	6 (*)	7 (*)	8 (*)	9 (*)
01	120 678	19,93	2,93	8,95	2,34	3,62	4,69	31,12	,81	2,80	22,82
02	176 941	7,06	9,28	6,61	2,47	4,19	5,25	36,78	,70	2,32	25,34
03	140 885	15,70	8,63	8,40	2,32	3,88	4,79	27,81	,91	1,73	25,84
04	36 063	19,03	4,74	9,03	2,62	3,48	3,74	28,23	,65	2,58	25,90
05	33 221	23,61	3,02	8,61	2,60	3,90	4,58	24,55	,91	3,60	24,63
06	236 442	4,49	2,26	12,00	4,27	4,60	6,15	27,74	3,80	3,03	31,67
07	91 282	25,20	4,61	7,67	1,72	3,07	3,72	26,70	,52	1,51	25,29
08	104 691	8,72	4,64	6,25	2,75	4,35	5,52	41,49	,53	2,43	23,30
09	52 826	23,45	4,33	8,92	1,89	2,97	3,31	24,07	,55	2,10	28,40
10	91 035	10,91	5,41	7,45	2,69	4,22	5,80	35,46	,78	2,39	24,89
11	102 039	16,35	15,54	8,85	2,04	3,34	4,09	19,06	,61	2,30	27,83
12	110 185	29,55	5,81	8,93	1,45	2,64	3,21	19,13	,42	1,66	27,19
13	463 851	3,14	2,50	7,78	5,04	5,39	8,09	35,21	2,10	3,16	27,60
14	160 565	12,30	8,40	7,90	2,93	4,54	5,04	32,33	,98	1,70	23,88
15	63 437	27,49	14,55	9,49	1,50	2,44	3,24	15,23	,47	1,17	24,42
16	118 386	19,37	8,61	8,28	2,15	3,21	4,51	25,71	,56	2,85	24,74
17	167 983	16,46	6,17	10,68	2,15	3,27	4,92	24,57	,73	3,75	27,29
18	108 934	12,49	8,41	7,56	2,33	4,00	4,49	30,60	,68	2,75	26,70
19	89 887	26,25	6,15	9,37	1,94	3,29	4,03	21,87	,55	1,36	25,22
20	69 440	14,60	7,43	7,86	1,93	3,43	3,95	15,06	1,04	4,61	40,09
21	138 366	11,01	4,98	7,28	3,57	5,26	5,69	31,06	,82	4,16	26,17
22	176 986	28,31	6,16	8,29	1,91	2,86	2,85	19,71	,58	2,59	26,73
23	62 862	33,68	10,69	9,03	1,42	2,33	3,03	15,12	,48	1,05	23,17
24	139 563	26,42	7,99	8,65	1,72	2,85	3,79	20,25	,59	2,04	25,70
25	135 275	10,07	1,50	6,26	3,10	5,79	5,13	41,38	,84	2,74	23,19
26	111 226	18,23	4,53	8,29	2,83	4,52	5,33	29,99	,66	1,95	23,68
27	124 985	10,66	9,51	7,88	2,62	3,90	4,49	35,51	1,05	2,19	22,19
28	98 922	11,75	8,75	7,49	2,55	4,06	5,25	31,90	,94	2,69	24,62
29	265 350	20,66	4,15	7,86	2,14	3,67	2,95	26,51	,46	4,73	26,85
30	158 240	10,60	7,71	7,37	1,90	4,06	4,97	29,35	,59	2,87	30,58
31	220 025	11,49	4,29	8,56	4,07	6,06	6,28	27,86	1,00	2,83	27,54
32	69 435	38,95	11,10	8,43	1,43	2,27	2,61	12,19	,39	1,56	21,08
33	334 217	8,60	7,04	9,49	3,83	5,04	6,34	29,96	1,08	3,03	25,59
34	188 614	10,78	11,71	8,64	3,15	4,25	5,50	22,45	,80	2,17	30,54
35	210 417	22,12	5,11	8,28	2,69	4,10	4,30	25,37	,62	2,56	24,85
36	92 645	19,28	11,01	8,59	2,35	3,22	4,90	23,05	,67	2,57	24,37
37	140 152	13,25	7,79	8,16	2,85	4,34	5,31	28,48	,93	2,81	26,08
38	267 172	10,20	1,94	7,71	3,71	5,27	4,58	41,76	,85	1,64	22,33
39	81 049	17,08	2,69	9,41	2,28	3,62	4,12	31,69	,54	2,48	26,09
40	96 052	26,03	9,80	8,29	2,07	2,64	3,33	22,51	,75	3,41	21,17
41	89 907	18,60	9,65	8,44	1,98	3,29	4,29	25,74	,78	2,04	25,19
42	250 013	9,20	2,04	7,97	2,94	4,82	4,92	41,71	,59	1,38	24,42
43	76 604	32,05	3,36	9,04	1,58	2,68	3,25	20,95	,42	1,38	25,29
44	270 473	13,77	4,07	7,69	3,01	5,49	4,75	34,68	,85	2,09	23,59
45	140 321	11,97	5,48	7,59	3,25	4,43	5,59	31,86	,95	4,20	24,70
46	55 656	33,68	5,41	9,24	1,65	2,87	3,51	16,59	,59	1,57	24,89
47	101 744	27,55	7,11	8,78	2,03	3,11	4,01	20,74	,54	1,97	24,17
48	31 015	30,95	8,06	7,50	1,58	2,98	3,18	16,34	,57	2,63	26,22
49	185 917	19,28	8,91	7,87	2,53	3,92	4,34	26,22	,70	2,20	24,03
50	149 732	26,61	8,45	8,27	2,02	3,13	3,65	22,74	,57	2,45	22,11
51	155 261	10,05	7,15	6,75	3,24	4,60	6,14	33,83	,77	3,10	24,37
52	72 516	13,28	4,60	7,31	2,57	3,98	5,02	35,49	,67	3,28	23,80
53	85 206	30,44	9,85	8,19	1,58	2,88	3,31	20,39	,47	1,63	21,27
54	240 026	3,66	1,26	5,31	3,81	5,60	6,35	45,97	,97	3,54	23,62
55	74 903	12,42	3,70	6,21	2,85	3,96	5,41	35,48	,60	4,89	24,48
56	184 818	24,60	5,53	7,88	1,93	3,14	2,69	25,40	,48	3,63	24,71
57	332 298	3,88	1,21	4,67	2,90	5,16	5,56	52,16	,90	3,19	20,37
58	89 816	13,63	7,99	7,70	2,09	3,74	4,15	29,78	,75	1,62	28,55
59	793 039	3,57	1,50	6,99	3,39	5,37	6,72	44,61	,78	1,90	25,18
60	166 419	5,43	6,04	6,77	2,80	4,62	4,92	41,24	1,13	2,25	24,79
61	95 396	22,29	10,29	8,18	1,93	3,31	4,26	25,56	,60	1,75	21,83
62	463 635	6,81	3,19	6,26	1,88	3,70	4,49	42,76	,57	1,59	28,75
63	187 399	16,78	3,15	8,42	2,89	4,34	4,65	33,20	,69	1,72	24,18
64	168 204	20,24	4,22	9,13	3,02	4,27	4,77	27,01	1,19	3,04	23,10
65	77 834	19,97	3,04	8,66	2,61	4,42	4,18	29,03	1,00	2,35	24,74
66	95 009	15,14	11,48	10,12	2,64	3,81	4,71	20,85	,94	2,93	27,39
67	272 838	9,09	1,77	6,65	4,04	5,64	7,54	36,96	,87	2,73	24,70
68	195 180	6,69	1,61	6,04	3,58	5,32	6,57	42,66	,84	2,58	24,10
69	406 174	4,24	1,30	8,16	5,51	7,40	7,40	39,95	1,31	2,00	22,74
70	112 366	13,34	1,99	6,97	2,67	4,78	4,88	36,53	,56	3,03	26,26
71	194 792	19,42	3,40	7,73	2,10	3,89	3,93	33,23	,60	1,45	24,26
72	148 828	17,89	6,94	7,44	2,22	4,10	5,23	29,99	,74	1,84	23,61
73	97 409	15,15	1,77	8,37	3,25	4,44	4,91	35,94	1,07	2,58	22,51
74	121 277	15,48	2,58	10,62	3,12	4,35	4,54	35,07	1,24	2,28	20,71
75	2 939 131	,40	,60	7,18	8,17	10,37	9,53	36,23	2,64	2,81	22,06
76	356 050	6,08	4,10	6,59	3,69	5,07	6,30	42,60	1,52	1,70	22,35
77	188 716	3,71	5,73	7,47	3,71	5,47	6,48	38,36	1,27	3,31	24,48
79	115 015	25,82	9,05	8,24	2,00	3,01	3,99	21,01	,51	2,24	24,14
80	169 093	10,55	7,64	7,75	2,34	4,32	5,40	33,23	,63	1,86	26,29
81	118 134	20,13	4,89	9,19	2,16	3,35	3,79	28,63	,56	1,64	25,65
82	64 094	28,80	7,88	8,59	1,81	2,76	3,42	19,75	,50	2,39	24,09
83	179 162	6,54	4,30	8,77	3,99	4,37	4,42	29,30	1,21	8,97	28,13
84	112 886	15,80	7,54	9,52	2,89	3,94	5,11	26,67	,89	2,65	24,99
85	139 526	27,20	9,00	10,24	1,40	2,56	2,99	21,53	,43	1,80	22,85
86	119 489	17,53	11,80	8,18	2,38	3,41	5,10	21,08	,72	2,70	27,10
87	124 256	17,32	7,38	8,72	2,46	4,02	5,17	27,14	,69	1,91	25,19
88	128 964	10,84	2,40	7,44	2,78	4,15	5,11	40,91	,65	2,26	23,45
89	98 645	13,95	7,21	8,57	2,02	3,48	4,26	26,84	,76	2,61	30,32
FRANCE ENTIÈRE	16 692 283	11,10	4,38	7,76	3,86	5,44	5,91	33,18	1,20	2,59	24,59

(*) 0 Agriculteurs exploitants.
1 Salariés agricoles.
2 Patrons de l'industrie et du commerce.
3 Professions libérales et cadres supérieurs.
4 Cadres moyens.
5 Employés.
6 Ouvriers.
7 Personnel de service.
8 Autres catégories.
9 Non-actifs.

CATEGORIES SOCIO-PROFESSIONNELLES MASCULINES EN 1968

Pourcentages par rapport à la population masculine de 15 ans et plus, active et non active

ANNEXE VI (C)

DEPARTEMENTS	POPULATION MASCULINE de 15 ans et plus	0 (*)	1 (*)	2 (*)	3 (*)	4 (*)	5 (*)	6 (*)	7 (*)	8 (*)	9 (*)
01	127 536	14,30	2,04	8,08	2,66	5,02	5,51	35,75	1,04	2,15	23,47
02	186 396	6,04	6,57	5,50	2,61	5,36	5,79	37,37	88	2,05	27,83
03	145 492	12,37	5,47	7,69	2,71	4,72	5,62	29,56	1,01	1,65	29,21
04	40 328	12,99	3,75	9,16	3,18	4,76	4,46	29,75	,84	2,39	28,70
05	35 812	18,11	1,76	8,79	2,96	4,78	5,62	25,83	1,56	3,25	27,33
06	277 204	2,98	1,63	11,26	4,70	5,67	6,83	28,59	3,61	2,30	32,45
07	96 420	17,90	2,83	7,65	1,96	4,23	4,81	30,11	,79	1,34	28,38
08	108 072	7,11	3,06	5,64	2,96	5,50	5,90	40,37	,64	2,27	26,54
09	53 924	16,55	2,79	7,86	2,31	4,01	4,12	25,52	,83	1,81	34,21
10	99 192	8,72	3,58	6,34	2,95	5,93	6,55	36,60	,97	2,10	26,27
11	105 592	12,89	11,44	8,23	2,53	4,19	4,83	21,15	,67	2,16	31,92
12	106 936	25,55	3,83	8,02	1,97	3,77	4,44	20,34	,55	1,36	30,18
13	548 884	2,32	1,81	6,94	5,17	6,49	7,92	33,62	2,04	2,95	30,75
14	178 732	9,70	5,85	6,93	3,41	6,23	5,53	34,38	1,14	1,55	25,26
15	64 032	24,58	9,21	8,74	1,78	3,17	4,17	18,53	,60	1,17	28,04
16	121 184	15,15	6,92	8,31	2,61	3,98	5,20	28,84	,67	1,78	26,55
17	177 040	13,29	4,36	9,95	2,51	4,22	5,21	27,65	,84	2,99	28,98
18	114 920	9,60	5,85	6,87	2,74	5,22	4,71	33,01	,85	2,35	28,77
19	91 708	21,06	3,53	8,73	2,40	4,56	4,75	24,30	,56	1,23	28,87
20	82 680	9,22	7,40	8,35	2,49	3,14	4,89	19,52	1,43	5,88	37,69
21	153 844	8,27	3,27	6,64	4,25	6,64	7,10	32,49	,97	2,87	27,50
22	181 888	22,28	4,03	7,89	2,48	3,98	4,02	23,20	,73	1,93	29,45
23	61 036	28,53	6,91	8,17	1,59	2,63	3,48	17,74	,43	1,08	29,43
24	142 940	20,62	5,15	8,62	2,14	3,46	4,10	23,20	,72	1,80	30,20
25	151 824	7,51	,93	5,77	3,61	7,17	6,08	41,36	,86	2,21	24,49
26	124 764	13,48	3,17	7,55	3,25	6,07	6,91	31,60	,90	2,21	24,86
27	136 700	8,37	6,23	7,30	2,82	5 05	5,10	39,27	1,21	1,47	23,19
28	110 480	9,04	4,91	6,78	2,89	5,68	5,50	36,75	1,12	2,15	25,17
29	276 580	15,35	2,62	7,70	2,78	4,47	4,18	28,15	,63	4,10	30,01
30	177 508	8,31	5,60	6,82	3,23	5,09	5,71	29,50	,80	3,01	31,95
31	260 416	7,69	2,38	7,34	5,10	7,43	7,40	28,08	1,07	2,39	31,13
32	70 256	31,79	7,16	7,69	1,83	3,19	3,50	16,04	,47	1,41	26,93
33	367 784	6,26	4,96	8,66	4,39	5,97	6,83	30,10	1,19	2,99	28,63
34	219 972	7,95	7,60	7,95	4,19	5,38	5,77	24,43	,98	1,92	33,85
35	228 460	16,95	3,17	7,13	3,36	5,31	5,47	29,07	,85	2,28	26,41
36	92 868	15,26	7,11	7,76	2,14	4,26	4,94	28,09	,91	1,55	28,00
37	158 452	9,63	4,96	7,34	3,59	5,96	5,50	32,38	1,12	2,51	27,01
38	280 008	7,18	1,11	7,01	4,69	7,01	5,39	39,28	1,10	1,56	25,66
39	85 580	13,06	1,80	8,53	2,49	4,65	5,45	33,67	,81	2,07	27,68
40	103 436	18,52	6,09	7,60	2,49	3,40	3,73	26,17	,83	4,65	26,52
41	98 532	13,53	6,18	7,94	2,14	4,75	4,79	31,22	1,03	1,93	26,48
42	264 920	6,90	1,30	7,15	3,10	6,10	5,35	40,92	,70	1,31	27,16
43	76 408	25,45	1,74	8,53	2,00	3,69	4,02	24,19	,46	1,28	28,64
44	298 008	10,63	2,61	6,86	3,63	6,80	5,70	34,65	1,01	1,69	26,42
45	159 428	8,50	3,63	6,89	3,79	6,22	5,52	35,61	1,08	2,61	26,14
46	56 528	26,27	3,24	8,63	2,10	3,97	4,42	19,37	,86	1,32	29,81
47	109 396	21,32	4,86	8,29	2,38	3,76	4,75	23,88	,73	1,82	28,21
48	29 704	27,58	5,22	7,31	2,14	4,42	3,85	16,36	,75	2,44	29,92
49	199 848	15,45	6,21	6,94	2,87	5,13	5,35	29,53	,84	1,76	25,92
50	156 060	21,93	6,04	7,38	2,20	4,35	4,29	26,80	,80	2,08	24,13
51	177 268	8,53	4,81	5,66	3,73	5,88	6,39	35,71	,86	3,13	25,31
52	77 468	11,20	3,18	5,94	2,63	4,84	5,83	36,11	,76	3,27	26,22
53	87 908	26,64	6,16	7,23	1,89	4,15	4,59	25,22	,70	1,63	21,79
54	255 236	2,86	1,08	4,73	4,31	6,56	6,91	41,69	,99	3,14	27,72
55	75 724	10 45	2,88	5,68	2,61	4,32	5,97	35,97	,66	3,52	27,93
56	191 568	18,35	3,34	7,19	2,55	4,21	3,80	29,58	,59	3,07	27,32
57	344 416	2,84	1,02	4,06	3,33	5,95	6,47	47,92	,98	2,63	24,79
58	92 360	11,08	5,61	7,42	2,31	4,69	5,67	29,68	,97	1,46	31,12
59	850 852	2,92	1,05	5,97	3,57	6,55	6,92	41,73	,93	1,68	28,69
60	194 120	4,29	4,84	5,86	3,12	6,37	5,45	41,57	1,23	2,47	24,81
61	100 876	18,41	7,38	7,30	2,19	4,64	4,82	29,89	,73	1,58	23,05
62	485 124	5,76	2,10	5,58	2,20	4,70	5,03	40,62	,72	1,53	31,76
63	208 240	12,03	1,88	7,63	3,85	5,39	5,22	34,93	,81	1,83	26,44
64	186 496	15,58	2,26	8,37	3,59	5,46	5,22	28,49	1,10	3,22	26,71
65	84 852	15,01	1,87	8,07	3,08	5,44	5,02	29,11	1,28	2,35	28,78
66	106 368	11,02	7,32	9,15	3,03	4,54	5,47	23,92	1,09	2,47	32,01
67	296 752	5,90	1,23	5,32	4,90	6,93	8,02	36,55	,99	2,47	27,69
68	209 316	4,77	1,15	4,96	4,27	6,68	6,97	42,21	1,01	2,54	25,46
69	481 304	3,37	,91	7,06	5,64	8,57	7,38	38,90	1,45	1,85	24,86
70	120 116	9,22	1,24	6,36	2,98	5,70	5,51	37,93	,69	2,92	27,45
71	203 732	14,45	2,37	7,12	2,45	4,89	4,98	34,46	,80	1,21	27,27
72	160 368	13,76	4,39	6,61	2,70	5,21	6,41	33,27	,94	1,64	25,08
73	108 032	10,30	1,00	7,81	3,48	5,48	5,36	37,99	1,31	2,92	24,35
74	138 396	9,86	1,25	9,95	3,63	6,17	5,60	38,86	1,22	1,86	21,60
75	3 255 328	,30	,47	6,22	9,23	11,12	9,31	34,29	2,73	2,49	23,84
76	392 704	4,86	2,65	5,80	4,06	6,14	6,55	43,18	1,42	1,60	23,73
77	222 968	2,70	3,49	6,97	4,33	6,98	6,85	40,01	1,55	2,36	24,77
79	118 244	21,42	5,00	7,72	2,17	3,88	5,64	25,67	,71	1,58	26,23
80	182 636	8,57	4,95	6,49	2,79	5,31	5,81	34,60	,82	1,67	29,01
81	124 768	14,88	2,94	8,26	2,29	4,41	4,48	29,94	,67	1,54	30,59
82	68 540	23,35	4,87	8,04	2,24	3,81	4,14	22,24	,81	2,01	28,49
83	210 428	4,96	3,25	8,74	4,01	5,41	5,11	30,07	1,54	6,26	30,66
84	131 224	12,10	5,23	8,35	3,23	5,41	5,69	30,10	1,08	2,75	26,07
85	147 300	21,63	4,22	9,49	1,85	3,60	3,79	26,86	,58	1,44	26,53
86	124 588	13,76	7,41	7,23	3,21	4,62	5,63	25,69	,95	2,14	29,36
87	130 900	13,13	4,52	7,68	3,10	5,23	5,56	29,34	,75	1,79	28,88
88	136 316	8,31	1,84	6,45	2,95	5,09	5,80	41,07	,85	2,15	25,49
89	106 192	10,72	4,44	7,61	2,55	4,64	5,35	30,28	,89	1,99	31,53
FRANCE ENTIÈRE	18 184 740	8,39	2,88	6,97	4,43	6,55	6,42	33,71	1,34	2,29	27,02

(*) 0 Agriculteurs exploitants.
1 Salariés agricoles
2 Patrons de l'industrie et du commerce.
3 Professions libérales et cadres supérieurs.
4 Cadres moyens.
5 Employés.
6 Ouvriers.
7 Personnel de service.
8 Autres catégories.
9 Non-actifs.

Imprimerie administrative

MELUN 3953-1971