Reuceil du Symposium 2001 de Statistique Canada
La qualité des données d'un organisme statistique : une perspective méthodologique

L'ESTIMATION POUR LE PROGRAMME AMÉRICAIN DE DÉCLARATION UNIFORME DE LA CRIMINALITÉ

Yoshio Akiyama[1] ,Samuel Berhanu[2]

RÉSUMÉ

Afin de combler les lacunes laissées par les données manquantes, non déclarées, déraisonnables ou inutilisables, le programme américain de déclaration uniforme de la criminalité procède à des estimations/ imputations faisant appel à plusieurs méthodes statistiques. Dans la présente communication, les auteurs abordent et illustrent l'estimation des données sur les infractions et les arrestations en utilisant diverses approches et en soulignant les forces et les faiblesses de chacune.

MOTS-CLÉS : Estimation des données; Données sur les infractions et les arrestations; Strates; Programme de déclaration uniforme de la criminalité (DUC).

1. INTRODUCTION

S'il recueille, stocke, traite et diffuse des données sur la criminalité, le programme américain de déclaration uniforme de la criminalité (programme DUC) procède aussi à une gamme de calculs statistiques simples et avancés afin d'offrir un aperçu des phénomènes nationaux, régionaux et locaux touchant la criminalité. Les principales publications annuelles du programme (*Crime in the United States, Law Enforcement Officers Killed and Assaulted*, etc.) et ses statistiques sur le crime motivé par la haine présentent des renseignements d'ordre statistique. Parmi les statistiques courantes présentées par le programme américain, on retrouve les variations en pourcentage, le taux de criminalité, les données sur l'horloge de la criminalité, le taux d'arrestation selon l'âge, le pourcentage des affaires classées et le pourcentage des employés de la police. Ces statistiques sont exprimées dans une langue simple et non technique, à la portée du grand public.

Pour les besoins des publications susmentionnées, on procède à l'estimation/imputation des données sur la criminalité lorsque les données déclarées par les organismes sont incomplètes ou inacceptables. La politique d'estimation/imputation des données sur la criminalité du programme DUC est en vigueur depuis 1960. Pour chaque organisme, l'estimation des données sur la criminalité est fondée sur la taille de l'organisme, le degré d'urbanisation, la région et le type d'organisme. La méthodologie de l'estimation est résumée à l'annexe I de la publication intitulée Crime in the United States. Cette méthodologie comporte les inconvénients suivants :

1. Un grand nombre de strates : Chaque État américain est stratifié selon huit groupes démographiques de DUC et par région statistique métropolitaine (RSM). Officiellement, cela donne lieu à 11 strates pour chaque État. Ce nombre est cependant trop élevé, il représente une stratification surperfectionnée et il donne lieu à un faible nombre d'organismes au sein d'une même strate. Le nombre d'organismes pour certaines strates s'est souvent avéré trop petit pour qu'on puisse estimer les données manquantes.

---

[1]Statisticien principal, Federal Bureau of Investigation, J. Edgar Hoover Building, Room 11194, 935 Pennsylvania Avenue, Washington, D.C., USA 20535.

[2] Statisticien, sous-section de l'analyse, de la recherche et du développement, division des services d'information sur la justice pénale, Federal Bureau of Investigation, 1000 Custer Hollow Road, Clarksburg, WV, USA 26306.

2. Un seuil arbitraire de trois mois de déclarations A utilisables : Les organismes sont classés d'après un seuil arbitraire de trois mois de déclarations utilisables. L'organisme de catégorie I fournit une série complète (douze mois) de déclarations utilisables pour tous les crimes de l'indice. L'organisme de catégorie II fait parvenir des déclarations utilisables sur moins de trois mois. Enfin, l'organisme de catégorie III fournit des déclarations A utilisables pour une période allant de trois à onze mois.

   Les données pour les organismes de catégorie II sont estimées par la méthode d'estimation par quotient (c'est-à-dire au moyen d'une répartition proportionnelle selon la population). Celles pour les organismes de catégorie III sont estimées au moyen d'une répartition proportionnelle mensuelle. La ventilation des organismes selon les catégories II et III est fondée sur le critère arbitraire de trois mois de déclarations utilisables.

3. La difficulté à calculer les erreurs-types : Du fait que les organismes de catégorie II et de catégorie III font l'objet d'une méthode d'estimation différente, il est fastidieux de calculer les erreurs-types. Il est particulièrement fastidieux d'associer la précision et l'exactitude aux estimations portant sur les organismes de catégorie III. Historiquement, les estimations de la précision (erreurs-types) n'ont pas été associées aux estimations de la DUC sur la criminalité.
4. La répartition proportionnelle mensuelle pour les organismes de catégorie III ne tient compte ni de la saisonnalité ni de la tendance : La répartition proportionnelle mensuelle pour les organismes de catégorie III ne tient pas compte des tendances et de la saisonnalité de la criminalité. Afin d'obtenir des estimations annuelles viables, il faut éliminer la tendance temporelle et la saisonnalité de ces répartitions proportionnelles.

Aussi la présente communication cherche-t-elle avant tout à exposer la nouvelle approche mise au point par le programme DUC afin d'estimer les données de la DUC.

2. STRATIFICATION DES ORGANISMES

Afin de relever les valeurs aberrantes, les algorithmes comparent la déclaration d'un organisme avec celles d'organismes semblables. La méthode privilégiée consiste à classer les organismes en strates. Les organismes appartenant à une même strate sont désignés des organismes semblables. La nouvelle méthode d'estimation est fondée sur les facteurs suivants :

- la taille du territoire (c.-à-d. la population desservie par l'organisme)
- la région où est situé l'organisme
- le degré d'urbanisation (c.-à-d. à l'intérieur ou à l'extérieur d'une RSM)
- le type d'organisme (c.-à-d. la distinction entre les services de police et les bureaux de shérif).

Les quatre facteurs ci-dessus servent à la stratification traditionnelle pour les besoins de la DUC (sauf les ventilations géographiques, présentées par État dans la stratification traditionnelle) et constituent l'ensemble *minimum* des variables ayant une incidence sur les conditions de criminalité du territoire. La ventilation géographique se fait par région (et non par État) afin d'accroître la taille des strates. Les gros organismes (population supérieure à 50 000) ne sont pas classés par région. Mises à part ces différences, la nouvelle stratification adopte les mêmes variables et la même notion que l'ancienne. Un plan de stratification sert à la fois au contrôle de la qualité des données (détection des valeurs aberrantes dans l'examen du caractère raisonnable des données) et à l'estimation des données (estimation des crimes et des arrestations).

La plupart des données publiées de la DUC sont complètes. Les données manquantes/incomplètes (représentées par les organismes des catégories II et III) sont l'exception plutôt que la règle. Le programme DUC dispose de données réelles pour la plupart des organismes. Comme on procède à des estimations pour un nombre relativement faible d'organismes et que le nombre total d'organismes est fini, un plan de stratification aurait

une incidence limitée sur les estimations d'ensemble. Considéré dans le cadre de l'échantillonnage, le *facteur de rajustement fini* (présent dans la variance de l'estimation de la DUC) ramène les erreurs d'estimation à près de zéro. Par comparaison, dans la plupart des enquêtes par sondage, les chercheurs doivent estimer les caractéristiques de la population à partir d'une *fraction de sondage* peu élevée.

3. NOUVELLES MÉTHODES D'ESTIMATION DES INFRACTIONS

3.1 Estimations par quotient des infractions

Selon la méthode traditionnelle d'estimation des infractions, le seuil fixé à trois mois de déclarations utilisables est arbitraire. On se souviendra que ce seuil de trois mois constitue la ligne de démarcation entre les organismes de catégorie II et de catégorie III, et que la répartition proportionnelle mensuelle a touché uniquement les organismes de catégorie III. Afin d'atténuer le problème découlant de ce seuil arbitraire, l'estimation des données sur les infractions se fait sur une base mensuelle.

Les organismes ayant une population positive et qui déclarent les données du programme DUC sur une base mensuelle n'ont guère de mal à faire estimer leurs données. Pour une strate donnée définie parmi les 32 strates (sauf celle à population nulle), supposons que n soit le nombre d'organismes dans la strate déclarant des données sur la criminalité du programme DUC sur une base mensuelle. Supposons ensuite qu'il existe j organismes dont les données sur la criminalité ne sont pas utilisables (c'est-à-dire qu'il faut les estimer), k organismes qui ont reçu la cote S (autoreprésentatifs) et $m = n - j - k$ organismes restants dont les données sont utilisables (c'est-à-dire qu'elles ne nécessitent aucune estimation). Dans ce contexte, un rapport mensuel est utilisable s'il n'exige pas d'estimation mensuelle et qu'il n'est pas coté S (autoreprésentatif).

Supposons que $X_i,\ 1 \# i \# m,$ soit le rapport mensuel utilisable du i^e organisme, et que P_i soit la population du i^e organisme de la strate. L'hypothèse veut qu'il y ait m organismes du genre. Supposons que :

$$X^{)} ' \sum_{i'1}^{m} X_i,\ P^{)} ' \sum_{i'1}^{m} P_i,\ \text{et}\ r ' \frac{X^{)}}{P^{)}}. \qquad [1]$$

Supposons que Q_i ($m\%1 \# i \# m\%j ' n\&k$) soit la population de l'organisme dont les données mensuelles sur la criminalité n'étaient pas utilisables (c'est-à-dire qu'il fallait les estimer) et que $Q ' \sum_{i} Q_i$. L'hypothèse veut qu'il y ait j organismes du genre. Le nombre de crimes estimés de l'organisme (nombre entier) pour le mois est définit par l'équation suivante :

$$\tilde{X}_i ' r\, Q_i,\ m\%1 \# i \# m\%j ' n\&k. \qquad [2]$$

Le nombre de crimes pour chacun des k organismes qui ont reçu la cote S (autoreprésentatifs) correspond au nombre réel $X_i^{(S)},\ n\&k\%1 \# i \# n,$ que les organismes ont déclaré. Supposons que :

$$\tilde{X} ' \sum_{i'm\%1}^{n\&k} \tilde{X}_i \quad \text{et} \quad X^{(S)} ' \sum_{i'n\&k\%1}^{n} X_i^{(S)}. \qquad [3]$$

Par conséquent, l'estimation totale de la criminalité pour le mois pour la strate correspond à l'équation suivante :

$T ' X^{)} \% \tilde{X} \% X^{(S)} \%$ (données réparties proportionnellement des organismes à population nulle). [4]

Pour une catégorie donnée de criminalité, l'estimation annuelle pour la strate s'obtient en additionnant les 12 estimations mensuelles T. L'estimation nationale pour le mois (et pour la catégorie de criminalité) s'obtient en additionnant les estimations mensuelles pour l'ensemble des strates. Étant donné qu'à chaque organisme

correspond le nombre réel X_i, le nombre estimé $\tilde{X}_i$ ou le nombre autoreprésentatif $X_i^{(S)}$, pour un mois donné, l'estimation de la criminalité peut s'obtenir pour une région donnée en additionnant les données mensuelles des organismes concernés.

3.2 Variances des estimations par quotient

Pour une strate donnée, la formule (4) définit l'estimation des infractions $T = rP + X^{(S)} + (\text{autres})$ pour une catégorie de la criminalité et un mois, lorsque $P = P' + Q$ correspond à la population totale de la strate, sans les organismes qui ont reçu la cote S (autoreprésentatifs). Le ratio r est une variable, car elle dépend de la catégorie des m organismes qui se trouvent à soumettre des données « utilisables » pour le mois. On sait que la formule suivante donne une approximation de la variance de r :

$$\sigma_r^2 \doteq R^2 \left(1 - \frac{m}{n-k}\right) \frac{V_X^2 + V_P^2 - 2\rho V_X V_P}{m}, \qquad [5]$$

où V dénote la *précision relative* et ρ correspond au coefficient de corrélation entre X_i et P_i. Comme j (le nombre d'organismes dont les données sont estimées) est habituellement peu élevé, le facteur de rajustement fini $1 - m/(n-k) = j/(n-k)$ dans l'équation ci-dessus est faible. Il convient de noter que les valeurs dans R, V_X^2, V_P^2, et ρ peuvent être estimées à partir des données provenant des m organismes (considérés comme un échantillon) qui sont à l'origine de déclarations « utilisables ». Dans ce cas, l'équation (5) se transforme ainsi :

$$\sigma_r^2 \doteq r^2 \left[\frac{s_X^2}{(X')^2} + \frac{s_P^2}{(P')^2} - 2\hat{\rho}\frac{s_X s_P}{X' P'}\right], \qquad [6]$$

où s_X^2, s_P^2, et $\hat{\rho}$ sont des estimations (pour l'échantillon) de σ_X^2, σ_P^2, et ρ.

Étant donné que $T = rP + X^{(S)} + (\text{autres})$ et que $X^{(S)}$ est une constante, la variance de T est donnée comme suit :

$$\sigma_T^2 = \sigma_r^2 P^2. \qquad [7]$$

En supposant l'indépendance des mois et des strates dans l'estimation des données, la variance combinée s'obtient en additionnant les variances appropriées.

3.3 Approche de l'estimation des données dite du filtre de Kalman

La technique du filtre de Kalman a été mise au point et utilisée, à l'origine, par des ingénieurs. Toutefois, elle sert à une grande diversité de fins dans d'autres secteurs tels que la prévision et le contrôle de la qualité. Nous proposons ci-dessous d'appliquer cette technique à l'estimation des données de la DUC dans une forme simplifiée (Meinhold, R. J. et Singpurwalla, 1983, 1989).

3.3.1 Modèle du filtre de Kalman

Au moment t, supposons que Y_t dénote la valeur *observée* d'une variable (pour l'application de la DUC, Y_t est le nombre déclaré de crimes pour un organisme) et que U_t soit l'état naturel *non observé*. U_t peut être

considéré comme le taux de criminalité réel, mais non observable, de l'organisme. Le rapport entre Y_t et U_t est spécifié par l'*équation dérivée des observations* suivante :

$$Y_t = U_t + \eta_t, \qquad [8]$$

où η_t représente une variable normale $N(0, V_t)$. Dans le modèle du filtre de Kalman, il est supposé que l'état naturel varie au fil du temps *t* et est contrôlé par l'*équation dérivée du système* suivante :

$$U_t = G_t U_{t-1} + \xi_t, \qquad [9]$$

où G_t est une quantité connue et ξ_t est une variable normale $N(0, W_t)$ qui est statistiquement indépendante de η_t.

Au moment $(t-1)$, ce qu'on sait de l'état naturel peut s'exprimer par la distribution normale *a posteriori*, comme suit :

$$U_{t-1} * \Omega_{t-1} \sim N(\hat{U}_{t-1}, \Sigma_{t-1}), \qquad [10]$$

où Ω_{t-1} représente l'ensemble des observations historiques $\{Y_0, Y_1, \ldots, Y_{t-1}\}$, $\hat{U}_{t-1}$ dénote la valeur prévue $E(U_{t-1} * \Omega_{t-1})$, et Σ_{t-1} dénote la variance $Var(U_{t-1} * \Omega_{t-1})$. Le processus s'amorce au moment $t = 0$ en choisissant correctement $\hat{U}_0$ et Σ_0.

Avant d'observer Y_t (qui n'est pas déclaré dans notre cas), la distribution *a priori* $U_t * \Omega_{t-1}$ provient de l'équation dérivée du système. Plus précisément, nous avons la moyenne et la variance suivantes :

$$E(U_t * \Omega_{t-1}) = G_t E(U_{t-1} * \Omega_{t-1}) + E(\xi_t) = G_t \hat{U}_{t-1}. \qquad [11]$$

$$Var(U_t * \Omega_{t-1}) = G_t^2 Var(U_{t-1} * \Omega_{t-1}) + Var(\xi_t) \qquad [12]$$

$$= G_t^2 \Sigma_{t-1} + W_t.$$

Par conséquent, la variable $U_t * \Omega_{t-1}$ est distribuée normalement comme ci-dessous.

$$U_t * \Omega_{t-1} \sim N(G_t \hat{U}_{t-1}, G_t^2 \Sigma_{t-1} + W_t). \qquad [13]$$

Si Y_t est observé ou déclaré, la distribution *a posteriori* $U_t * \Omega_t$ s'exprime comme suit :

$$U_t * \Omega_t \sim N(\hat{U}_t, \Sigma_t). \qquad [14]$$

Le théorème suivant fournit l'algorithme pour le calcul de la moyenne $\hat{U}_t$ et la variance Σ_t :

<u>Théorème</u> :

$$\hat{U}_t = G_t \hat{U}_{t-1} + \frac{G_t^2 \Sigma_{t-1} + W_t}{V_t + G_t^2 \Sigma_{t-1} + W_t} (Y_t - G_t \hat{U}_{t-1}). \qquad [15]$$

$$\Sigma_t = (G_t^2 \Sigma_{t-1} + W_t) - \frac{(G_t^2 \Sigma_{t-1} + W_t)^2}{V_t + G_t^2 \Sigma_{t-1} + W_t}. \qquad [16]$$

3.3.2 Application à la DUC du concept du filtre de Kalman

Le filtre de Kalman sert lorsqu'aucune donnée n'est déclarée pendant toute l'année par un État donné. Dans ces cas, on estime que les donnés historiques (longitudinales) de l'État fournissent une meilleure estimation que l'estimation transversale du quotient. Il est supposé que Y_t n'est pas déclaré dans le contexte actuel. Premièrement, nous devons définir $\hat{U}_0$ et Σ_0. G_t est défini comme le taux de variation (entre les années t & 1 et t) du volume de crimes. On peut obtenir cette valeur auprès des organismes qui ont présenté des données utilisables pendant des années consécutives dans la strate à laquelle l'organisme en question appartient. Il est supposé que η_t . $N(0, \hat{U}_t)$, c'est-à-dire où U_t sert de variable de Poisson dont le paramètre λ ' $\hat{U}_t$. Par conséquent, V_0 ' Y_0 et V_t ' $\hat{U}_t$.

W_t est défini comme ci-dessous. U_t obéissait par hypothèse à la loi de Poisson avec un paramètre λ ' $\hat{U}_t$, comme d'ailleurs la variable $U_t^{)}$ ' $G_t\, U_{t\&1}$ (car elle a également trait au nombre de crimes pour l'année *t*). L'écart $(U_t \& U_t^{)})$ des deux variables explicatives de Poisson a une variance $(2\,\hat{U}_t)$. Par conséquent, il est défini que W_t ' $2\,\hat{U}_t$. On peut maintenant calculer la distribution a priori $U_t * \cup_{t\&1}$ de l'état naturel.

Comme nous ne disposons pas d'un rapport utilisable Y_t, nous l'estimons au moyen de la formule $\hat{Y}_t$ ' $E(U_t * \cup_{t\&1})$ ' $\hat{U}_t$ ' $G_t\,\hat{U}_{t\&1}$, qui est l'équivalent de $\hat{Y}_t$ ' $(G_t\, G_{t\&1} \ldots G_1)\, Y_0$. Cette estimation de sens commun $\hat{Y}_t$ ' $(G_t\, G_{t\&1} \ldots G_1)\, Y_0$ est prévue dès le départ sans qu'on fasse appel au modèle du filtre de Kalman. Toutefois, le modèle du filtre de Kalman fournit la variance des estimations et fait appel à la déclaration la plus récente de l'organisme (au lieu d'appliquer la moyenne des organismes semblables).

4. NOUVELLE MÉTHODE D'ESTIMATION DES ARRESTATIONS

4.1 Estimation du nombre des arrestations

La nouvelle méthode d'estimation des arrestations convient à chaque catégorie de crime. S'agissant des algorithmes d'estimation, l'estimation par quotient s'applique aux déclarations utilisables des arrestations. La stratification des organismes se fait de la même façon que ci-dessus. Pour simplifier les choses, une catégorie d'infraction donnée est supposée ci-dessous lorsqu'on décrit la nouvelle méthode.
Dans la i^e strate et pour une catégorie d'infraction donnée, supposons ce qui suit :

n_{ij} = le nombre annuel d'arrestations déclarées par le j^e organisme utilisable;
p_{ij} = la population du j^e organisme utilisable pour les besoins de la DUC;
s_{ik} = le nombre annuel total des arrestations déclarées par le k^e organisme autoreprésentatif ayant reçu la cote S;
q_{ik} = la population du k^e organisme autoreprésentatif;
r_{im} = la population du m^e organisme pour lequel il faut estimer le nombre des arrestations.

Supposons que n_i ' $\sum_j n_{ij}$, p_i ' $\sum_k p_{ij}$, s_i ' $\sum_k s_{ik}$, q_i ' $\sum_m q_{ik}$, et r_i ' $\sum_m r_{im}$.

Le nombre estimé des arrestations pour le m^e organisme (pour lequel il faut estimer les données sur les arrestations) est défini par l'équation suivante :

$$\tilde{n}_{im} \text{ ' } \frac{n_i}{p_i}\, r_{im}. \qquad [17]$$

Il convient de noter que l'équation [17] est fondée sur l'estimation par quotient (échantillon) du nombre des arrestations et de la population; l'échantillon correspond à l'ensemble des organismes utilisables pour la catégorie de criminalité en cause. La partie estimée des arrestations pour la i^e strate est la suivante :

$$\tilde{n}_i \text{ ' } \sum_m \tilde{n}_{im}\,. \qquad [18]$$

Les nombres restants n_i et s_i sont des nombres réels. Par conséquent, le nombre total d'arrestations pour la i^e strate, $\tilde{N}_i$, représente la somme de n_i, s_i, et $\tilde{n}_i$, c'est-à-dire :

$$\tilde{N}_i = n_i + s_i + \tilde{n}_i, \qquad [19]$$

Aucune estimation n'est faite pour les organismes à population nulle (la strate 33), si bien que $\tilde{N}_{33}$ représente la somme des chiffres effectivement soumis par les organismes à population nulle. Par conséquent, l'estimation nationale des arrestations $\tilde{N}$ pour la catégorie d'infraction prescrite est la somme de N_i's :

$$\tilde{N} = \sum_{i=1}^{33} \tilde{N}_i. \qquad [20]$$

4.2 Estimation des arrestations selon l'âge et le sexe

Pour une catégorie d'infraction donnée, supposons que $n_{ij}(\alpha, \beta)$ correspondent au nombre de personnes arrêtées âgées de α ans et de sexe β. Le nombre $n_{ij}(\alpha, \beta)$ découle de la catégorie des organismes utilisables dans la i^e strate. Si l'on utilise l'indice *j* pour les organismes * utilisables + comme ci-dessus, nous obtenons :

$$n_{ij} = \sum_{\alpha} \sum_{\beta} n_{ij}(\alpha, \beta). \qquad [21]$$

Supposons que $n_i(\alpha, \beta) = \sum_j n_{ij}(\alpha, \beta)$. Puis, $n_i = \sum_{\alpha, \beta} n_i(\alpha, \beta)$.

Parallèlement, supposons que $s_{i,k}(\alpha, \beta)$ corresponde au nombre de personnes arrêtées de sexe β, âgées de α ans pour les organismes autoreprésentatifs et que $s_i(\alpha, \beta) = \sum_k s_{ik}(\alpha, \beta)$. (Les organismes autoreprésentatifs déclarent les données sur l'âge et le sexe pour les personnes arrêtées.) Par conséquent,

$$s_{ik} = \sum_{\alpha} \sum_{\beta} s_{ik}(\alpha, \beta). \qquad [22]$$

si l'on dénote $s_i(\alpha, \beta) = \sum_k s_{ik}(\alpha, \beta)$, puis $s_i = \sum_{\alpha, \beta} s_i(\alpha, \beta)$.

Pour l'organisme *m* dans la i^e strate (dont on a estimé les données sur les arrestations), le nombre estimé des arrestations de personnes de sexe β et âgées de α ans se définit par l'équation suivante :

$$\tilde{n}_{im}(\alpha, \beta) = \frac{n_i(\alpha, \beta)}{n_i} \tilde{n}_{im} = \frac{n_i(\alpha, \beta)}{p_i} r_{im}.$$

L'estimation pour la i^e strate est donc la suivante :

$$\tilde{N}_i(\alpha, \beta) = n_i(\alpha, \beta) + s_i(\alpha, \beta) + \frac{n_i(\alpha, \beta)}{n_i} \tilde{n}_i, \qquad [23]$$

La discussion ci-dessus exclut la strate 33 des organismes à population nulle. Afin d'inclure la strate 33, il faut tenir compte des définitions suivantes. Supposons que l'estimation nationale des arrestations pour les personnes d'âge α et de sexe β (sans la strate 33) soit définie par l'équation suivante :

$$\hat{N}(\alpha, \beta) = \sum_{i=1}^{32} \tilde{N}_i(\alpha, \beta) \text{ et} \qquad [24]$$

$$\hat{N} = \sum_{\alpha} \sum_{\beta} \hat{N}(\alpha, \beta).$$

Pour l'organisme *z* de la strate 33, supposons que $n_{33,z}(\alpha, \beta)$ soit le nombre réel de personnes arrêtées de sexe β, âgées de α ans. Ces données peuvent ne pas contenir les statistiques exhaustives des arrestations de l'année.

L'estimation pour le m^{e} organisme se définit par l'équation $\tilde{n}_{33,z}(\alpha,\beta) = n_{33,z}(\alpha,\beta)$, lorsque l'organisme *z* a déclaré les données sur l'âge et le sexe. Autrement, l'estimation se définit comme suit :

$$\tilde{n}_{33,z}(\alpha,\beta) = \frac{\hat{N}(\alpha,\beta)}{\hat{N}}\, n_{33,z}.$$

Maintenant, à partir de la notation $\tilde{N}_{33}(\alpha,\beta) = \sum_{z} \tilde{n}_{33,z}(\alpha,\beta)$, nous définissons l'estimation nationale des arrestations pour les personnes de sexe β âgées de α ans par l'équation suivante :

$$\tilde{N}(\alpha,\beta) = \sum_{i=1}^{33} \tilde{N}_i(\alpha,\beta) = \hat{N}(\alpha,\beta) + \tilde{N}_{33}(\alpha,\beta).$$

Il convient de noter que :

$$\tilde{N}_i = \sum_{\alpha}\sum_{\beta} \tilde{N}_i(\alpha,\beta), \text{ et} \qquad [25]$$

$$\tilde{N} = \sum_{\alpha}\sum_{\beta} \tilde{N}(\alpha,\beta).$$

4.3 Estimation des arrestations selon la race

Une méthode un peu différente sert à l'estimation du nombre des arrestations pour une race donnée γ. L'estimation de la race des personnes arrêtées est fondée sur les renseignements fournis par les organismes utilisables qui ont également déclaré de l'information sur la race des personnes arrêtées. Supposons que l'organisme utilisable (dans la i^{e} strate) qui a déclaré l'information sur la race des personnes arrêtées soit indexé par $\hat{j}$. Supposons que $n_{i\hat{j}}(\gamma)$ corresponde au nombre de personnes arrêtées de race γ, et que $p_{i\hat{j}}$ soit la population de l'organisme. Supposons en outre que $n_i(\gamma) = \sum_{\hat{j}} n_{i\hat{j}}(\gamma)$, $n_i^{(r)} = \sum_{\gamma} n_i(\gamma)$, et $p_i^{(r)} = \sum_{\hat{j}} p_{i\hat{j}}$. L'exposant (r) rappelle que le chiffre a trait à l'information sur la race. Nous avons les inégalités suivantes : $n_i^{(r)} \le n_i$ et $p_i^{(r)} \le p_i$.

Supposons que l'indice $\hat{k}$ dénote l'organisme autoreprésentatif qui a déclaré l'information sur la race des personnes arrêtées. Pour l'organisme $\hat{k}$, supposons que $s_{i\hat{k}}(\gamma)$ dénote le nombre de personnes arrêtées de race γ et que $q_{i\hat{k}}$ soit la population correspondante. Supposons enfin que $s_i(\gamma) = \sum_{\hat{k}} s_{i\hat{k}}(\gamma)$, $s_i^{(r)} = \sum_{\gamma} s_i(\gamma)$, et $q_i^{(r)} = \sum_{\hat{k}} q_{i\hat{k}}$. Manifestement, $s_i^{(r)} \le s_i$ et $q_i^{(r)} \le q_i$.

La population couverte par l'information sur la race des personnes arrêtées correspond à $p_i^{(r)} + q_i^{(r)}$. Par conséquent, la population non couverte par l'information sur la race est donnée par l'équation $u_i^{(r)} = p_i + q_i + r_i - p_i^{(r)} - q_i^{(r)}$. L'estimation du nombre d'arrestations pour la catégorie raciale γ de l'organisme *t* qui n'a pas déclaré les renseignements sur la race des personnes arrêtées se définit comme suit :

$$\tilde{n}_{it}(\gamma) = \frac{n_i(\gamma)}{n_i^{(r)}}\, n_{it}, \qquad [26]$$

si t ($= j$) représente un organisme admissible qui n'a pas déclaré la race des personnes arrêtées.

$$\tilde{n}_{it}(\gamma) = \frac{n_i(\gamma)}{n_i^{(r)}}\, s_{it}, \qquad [27]$$

si t (' k) représente un organisme autoreprésentatif qui n'a pas déclaré la race des personnes arrêtées.

$$\tilde{n}_{it}(\gamma) \text{ ' } \frac{n_i(\gamma)}{n_i^{(r)}} \tilde{n}_{it}, \qquad [28]$$

si t (' m) représente un organisme pour lequel il a fallu estimer le nombre des arrestations.

Évidemment, $\tilde{n}_{it}(\gamma)$ ' $n_{ij}(\gamma)$ si t (' j) représente un organisme utilisable qui a déclaré la race des personnes arrêtées. Pareillement, $\tilde{n}_{it}(\gamma)$ ' $s_{it}(\gamma)$ si t (' k) est un organisme autoreprésentatif qui a déclaré l'information sur la race des personnes arrêtées. Lorsque le total est calculé pour les γ, $\tilde{n}_{it}(\gamma)$ correspond au nombre réel ou estimé des arrestations de l'organisme obtenu selon la nouvelle méthode d'estimation des arrestations. Pour la i^{e} strate, le total estimé des arrestations pour la catégorie raciale γ est donné par l'équation suivante :

$$\tilde{N}_i(\gamma) \text{ ' } \sum_t \tilde{n}_{it}(\gamma) \text{ ' } n_i(\gamma) \% s_i(\gamma) \% \frac{n_i(\gamma)}{n_i^{(r)}} [\tilde{N}_i \& n_i^{(r)} \& s_i^{(r)}], \qquad [29]$$

lorsque $n_i^{(r)}$ ' $\sum_\gamma n_i(\gamma)$, $s_i^{(r)}$ ' $\sum_\gamma s_i(\gamma)$, et i ' $1, 2, \ldots, 32$. Par conséquent, nous obtenons :

$$\sum_\gamma \tilde{N}_i(\gamma) \text{ ' } \tilde{N}_i. \qquad [30]$$

Jusqu'ici, les organismes de la strate 33 (organismes à population nulle) n'ont pas été pris en compte. Aux fins du calcul des estimations raciales pour les organismes de la strate 33, on fait appel aux définitions suivantes :

$$\hat{N}(\gamma) \text{ ' } \sum_{i}^{32}{}_1 \tilde{N}_i(\gamma), \quad \hat{N} \text{ ' } \sum_\gamma \hat{N}(\gamma).$$

Pour l'organisme z de la strate 33, nous définissons $\tilde{n}_{33,z}$ ' $n_{33,z}$ lorsque l'organisme a déclaré les renseignements sur la race des personnes arrêtées. Autrement, nous définissons :

$$\tilde{n}_{33,z}(\gamma) \text{ ' } \frac{\hat{N}(\gamma)}{\hat{N}} n_{33,z}.$$

Au moyen de la notation $\tilde{N}_{33}(\gamma)$ ' $\sum_z \tilde{n}_{33,z}(\gamma)$, nous définissons $\tilde{N}(\gamma)$ ' $\sum_{i}^{33}{}_1 \tilde{N}_i(\gamma)$. Puis,

$$\tilde{N}_i \text{ ' } \sum_\gamma \tilde{N}_i(\gamma), \text{ pour } i \text{ ' } 1, 2, \ldots, 33, \qquad [31]$$

et, par conséquent,

$$\tilde{N} \text{ ' } \sum_\gamma \tilde{N}(\gamma).$$

La nouvelle méthode d'estimation des arrestations diffère des méthodes traditionnelles. Elle est fondée sur une stratification plus raffinée et considère à part les organismes autoreprésentatifs. Les nouvelles estimations portant expressément sur l'âge et sur la race sont fondées sur des strates multiples plutôt que sur une seule strate et définissent les estimations compatibles (dans la mesure où, au niveau de l'organisme, la ventilation selon l'âge et le sexe et celle selon la race s'additionnent pour donner le nombre réel ou estimé des arrestations) des arrestations selon l'âge, le sexe et la race pour chaque organisme.

5. CONCLUSION

Le programme DUC américain a fait l'objet d'un processus dynamique d'amélioration des méthodes servant à l'estimation des données sur les infractions et les arrestations. Le processus donne des estimations et assure également le niveau voulu de précision statistique. La nouvelle méthode constitue à bien des égards un pas en avant par rapport à la méthode traditionnelle. Le résultat obtenu en faisant appel à cette démarche est de loin supérieur à celui que donne l'approche traditionnelle. Ces améliorations sous-entendent qu'on a sensiblement accru la qualité des données de la DUC.

BIBLIOGRAPHIE

Meinhold, R. J. et Singpurwalla, (1983), AUnderstanding the Kalman Filter@, The American Statistician, Vol. 37, No.2., pp. 123-127.

Meinhold, R. J. et Singpurwalla, (1989), ARobustification of Kalman Filter Models@, Journal of the American Statistical Association, June 1989, Vol. 84, No. 406, pp. 479-486