# Projet VOIESUR - Livrable L3 - Méthodologie : redressement et extrapolation, projet VOIESUR - ANR

Emmanuelle Amoros, Audrey Lardy, Dan Wu, Vivian Viallon, Jean-Louis Martin

▶ **To cite this version:**

Emmanuelle Amoros, Audrey Lardy, Dan Wu, Vivian Viallon, Jean-Louis Martin. Projet VOIESUR - Livrable L3 - Méthodologie : redressement et extrapolation, projet VOIESUR - ANR. [Rapport de recherche] IFSTTAR - Institut Français des Sciences et Technologies des Transports, de l'Aménagement et des Réseaux. 2015, 63 p. hal-01212490v2

**HAL Id: hal-01212490**
**https://hal.science/hal-01212490v2**

Submitted on 2 Dec 2015

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

| | |
|---|---|
| **V**éhicule **O**ccupant **I**nfrastructure **E**tudes de la **S**écurité des **U**sagers de la **R**oute | Programme Transports Terrestres Durables Edition 2011 |

**Référence ANR du projet : ANR-11-VPTT-0007**

**Livrable L3**

# Méthodologie redressement et extrapolation

| | |
|---|---|
| **Date contractuelle de livraison du rapport :** | **juin 2013** |
| **Date de livraison du rapport :** | **avril 2015 = version 2.1** |
| **Auteur(s)** | **Emmanuelle Amoros, Audrey Lardy, Dan Wu, Vivian Viallon, Jean-Louis Martin** |
| **Participant(s)** | |
| **Tâche 3** | |
| **Niveau de confidentialité = Version 2.1** | **aucun ; public** |

**Résumé** :
Le projet VOIESUR se base sur l'informatisation la plus complète possible des Procès Verbaux (PV) dressés par les forces de l'ordre pour chaque accident de la route corporel qui leur est signalé. L'échantillonnage a consisté, pour l'année 2011, à inclure l'ensemble des PV mortels, et 1/20$^{ème}$ des PV corporels, ainsi que l'ensemble des PV du département du Rhône. L'exploitation statistique de ces données nécessite de tenir compte de cet échantillonnage dès que les accidents corporels et mortels sont analysés ensemble. Hors, si on peut supposer que tous les accidents mortels font l'objet d'un PV, il n'en est pas de même des accidents corporels dont le recueil n'est pas exhaustif, et biaisé sur de nombreux facteurs. De plus les statistiques officielles sont basées sur l'exploitation des BAAC qui sont un résumé informatique des PV par les forces de l'ordre, mais la correspondance PV -BAAC n'est pas toujours avérée.
L'objectif de cette tâche est de fournir les facteurs de pondération permettant de travailler sur l'ensemble des accidents corporels et d'extrapoler les résultats à l'ensemble de la France métropolitaine pour l'année 2011.
Les PV informatisés de VOIESUR sont dans un premier temps comparés à l'ensemble des BAAC d'accidents corporels. Comme ils différent, notamment sur la répartition des forces de l'ordre, du type d'usager, d'un tiers impliqué (oui/non), des coefficients de redressement $R_l$ estimés par post-stratification sont fournis pour redresser l'échantillon. Pour tenir compte du sous enregistrement biaisé des blessés par les forces de l'ordre, des coefficients de correction $C_j$ sont ensuite estimés par méthode de capture-recapture pour les années 2006 à 2012, en se basant sur la co-existence dans le Rhône des BAAC et du Registre médical du Rhône. Leurs valeurs dépendent de la gravité de l'accident (mortel /corporel), du type de force de l'ordre, du type de route, du type d'accident (avec ou sans tiers), du type d'usager (automobiliste/usager de 2RM/cycliste/piéton/autre), de la gravité du blessé (hospitalisé ou non), et de l'année.
Ces estimations permettent de proposer une pondération égale au produit du poids de sondage (1 pour les accidents mortels, 20 pour les accidents corporels), du coefficient de redressement $R_l$ (des PV vers les BAAC) et du coefficient de correction $C_j$, corrigeant des biais d'enregistrement. Dans les analyses statistiques, cette variable poids doit être déclarée dans les procédures d'estimations pour fournir les estimations de variance tenant compte des données réellement observées.
L'utilisation de ces pondérations permet d'obtenir des résultats représentatifs de la réalité de l'accidentalité sur la France entière à partir des données du projet.

**Mots clés** : blessés de la route ; données policières ; registre médical ; tirage aléatoire ; sous-enregistrement, biais, redressement ; prédiction ; capture-recapture ; projection ; standardisation indirecte ;

# 1 INTRODUCTION

Le projet VOIESUR se base sur l'exploitation des procès-verbaux (PV), documents papier établis par les forces de l'ordre pour les accidents corporels ou mortels de la route. Ces procès-verbaux font l'objet d'un enregistrement informatique sous forme de Bulletin d'Analyse d'Accident Corporel de la circulation routière (BAAC). Cet enregistrement est partiel, et sont notamment exclus : le récit en texte libre de l'accident, les auditions, les plans du lieu d'accident, les photos du lieu et des véhicules accidentés. C'est la richesse de ces informations, collectées dans les Procès-Verbaux, qui sont particulièrement exploitées dans VOIESUR.

Cependant, ce sont les BAAC qui font l'objet d'analyses régulières, et servent de bases aux communications officielles du ministère de l'intérieur et aux connaissances accidentologiques.
En théorie, à chaque procès-verbal vérifiant la définition d'un accident de la circulation correspond un enregistrement BAAC et un seul, et réciproquement. En pratique, ce n'est pas toujours le cas ; il arrive qu'un procès-verbal ait échappé à l'enregistrement BAAC, notamment si la procédure judicaire est longue, et inversement, il arrive que pour un enregistrement BAAC, le procès-verbal correspondant ait été égaré. Il importe donc, comme nous travaillons avec les procès-verbaux (documents papier), de vérifier qu'ils ne soient pas une distorsion des BAAC.

Dans le projet VOIESUR, l'ensemble des PV mortels sont collectés. En ce qui concerne les tués eux-mêmes, la déclaration et l'enregistrement par les forces de l'ordre sont bons : proche de l'exhaustivité et donc de la représentativité. Pour ceux-ci, il n'y a donc pas lieu de mettre en place une procédure de redressement ou de correction du sous-enregistrement. En revanche, pour les blessés dans des accidents mortels, il s'est avéré qu'il faut les redresser et les corriger du sous-enregistrement, tout comme les blessés dans des accidents corporels, mais avec des coefficients différents.

Pour les PV d'accidents corporels, seuls 1/20$^{ème}$ de ces PV sont tirés au sort et codés dans le projet VOIESUR (et la totalité pour le département du Rhône). Par rapport aux BAAC, cet échantillon est au 1/20$^{ème}$, et surtout il s'agit des Procès-verbaux (papier) et non des BAAC. C'est dans la transformation d'un PV en BAAC (par saisie informatique) qu'il pourrait y avoir des manques, et qu'il faut donc vérifier si l'échantillon des PV correspond bien aux BAAC (de façon globale et pas seulement sur les PV et BAAC effectivement reliés).
Des différences existent en fait, et ont montré la nécessité de redresser les blessés des PV. Des coefficients de redressement (qu'on note $R_l$) ont été estimés par post-stratification. Cette première étape de redressement par rapport aux BAAC est d'autant plus nécessaire, que la procédure de correction des biais de sous-enregistrement qui suit, est basée sur les BAAC.

En effet, et c'est le plus important, les données des forces de l'ordre sur les accidents corporels souffrent de sous-déclaration, sous-enregistrement et de distorsions associées à ce sous-enregistrement, dénommées « biais de sélection ». C'est-à-dire que certains types d'accidents corporels sont mieux enregistrés par les forces de l'ordre que d'autres accidents : ceux qui impliquent un tiers, ceux qui sont plus graves. A l'inverse, les accidents impliquant un deux-roues motorisé ou un vélo sont moins enregistrés que les accidents impliquant une voiture, ou un véhicule utilitaire. La probabilité d'enregistrement dépend aussi du type de réseau routier (autoroute, RN, RD, etc.), et du type de force de l'ordre (CRS / gendarmerie / police) (Amoros, Martin et al. 2006).

Une procédure d'extrapolation pour corriger ce sous-enregistrement, de son ampleur et surtout des biais associés, a déjà été mise en place sur la période 1996-2004 (Amoros, Martin et al. 2008). Cette procédure est mise à jour sur la période 2006-2012 (paragraphe 2), suite au changement en 2005 de

la définition des blessés graves dans les BAAC, et pouvant ainsi s'appliquer aux données VOIESUR de 2011. L'année 2005 est exclue des analyses car elle apparait comme une année de transition.

Cette procédure d'extrapolation recouvre une approche capture-recapture, suivie d'une projection, les deux étant basées sur certaines conditions. Cette procédure d'extrapolation permet indirectement d'estimer des coefficients de correction du sous-enregistrement des BAAC et des biais associés, qui, appliqués sur les BAAC donnent une estimation nationale représentative de la réalité de la morbidité routière en France.

Les coefficients de correction du sous-enregistrement estimés pour l'année 2011 seront appliqués, comme des poids, aux blessés des PV, déjà redressés. Ces coefficients (Cj) dépendent des caractéristiques qui jouent sur la probabilité de déclaration et d'enregistrement par les forces de l'ordre. Ainsi ils dépendent essentiellement de la gravité du blessé (hospitalisé / non-hospitalisé), de la gravité de l'accident (mortel /corporel), du type de force de l'ordre (CRS, gendarmerie, police), du type de réseau (autoroute, route nationale, départementale, communale ou autre), du type d'accident (=avec ou sans tiers), du type d'usager (piéton, cycliste, usager de 2RM, automobiliste, autre). En pratique cela représente 267 valeurs différentes. Chaque observation (un blessé dans un PV) se verra attribuer un coefficient correcteur en fonction des caractéristiques ci-dessus.

Les observations, que sont les blessés identifiés dans les PV se verront ainsi appliquer deux niveaux de poids : des coefficients Rl venant de la procédure de redressement, et des coefficients de correction Cj venant de la procédure d'extrapolation. Cela est résumé dans le schéma ci-après.

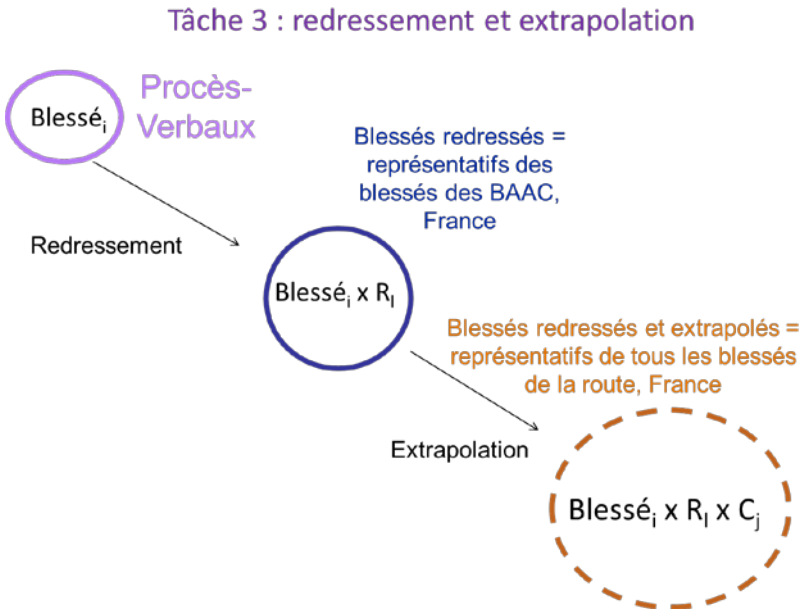


*Figure 1: représentation de la tâche 3 : redressement puis extrapolation des blessés des PV de VOIESUR*

De plus, les blessés dans des accidents corporels se verront attribuer un poids de 20, qui correspond au tirage au 1/20$^{ème}$, alors que les blessés dans les accidents mortels auront un poids de sondage de 1) ; nous notons ce poids pk.

L'ensemble de ces poids seront à prendre en compte lors des analyses spécifiques prévues dans la tâche 4.

# 2 REDRESSEMENT DES PV VOIESUR PAR RAPPORT AUX BAAC

## 2.1 Comparaison des PV de VOIESUR et des BAAC nationaux

Nous comparons, au niveau national, les PV de VOIESUR et les BAAC, pour l'année 2011. Il s'agit de vérifier, pour certaines variables, si la répartition dans les PV (au 1/20ème pour les accidents corporels, et la totalité pour les accidents mortels) est la même que celle des BAAC de la même année (il est à noter que nous ne nous restreignons pas aux PV et BAAC effectivement reliés).

Les variables que nous considérons sont celles qui pourraient jouer sur la transformation d'un PV en enregistrement BAAC, et celles qui servent à définir des sous-populations d'études dans la tâche 4.
Pour les premières, nous considérons celles qui jouent sur l'enregistrement final (en BAAC, comparé au Registre, cf. partie 3) d'un accident ou d'un blessé par les forces de l'ordre, et qui sont (cf. partie 3) la gravité de l'accident (mortel/ corporel), le type de force de l'ordre (CRS / Gendarmerie / Police), le type de réseau (autoroute (AR) / RN-RD / VC / autre), la gravité du blessé (hospitalisé ou non), le type d'usager (piéton / cycliste / usager de 2RM / automobiliste / autre), et la présence /absence d'un tiers.
Celles qui servent à définir des sous-populations d'étude dans la tâche 4 sont le type d'usager (études piétons, 2RM), et l'âge (études enfants, et séniors).
Cependant, la gravité de l'usager étant renseignée de façon experte dans VOIESUR, et pouvant donc différer de la gravité renseignée dans les BAAC, nous ne redresserons pas sur cette variable, quel que soit le résultat de la comparaison (dit autrement, si nous redressions sur cette variable, nous perdrions le bénéfice d'un codage expert).

Nous comparons, pour chaque variable citée, la répartition observée sur l'échantillon VOIESUR et la répartition des BAAC dans leur totalité en considérant cette dernière comme une répartition théorique, et en faisant ensuite le test de Chi-2 correspondant.

### 2.1.1 Comparaison PV VOIESUR – BAAC pour les accidents mortels, en termes de tués et de blessés

Il a été vérifié pour les tués, qu'il n'y avait pas de différence significative entre les tués de VOIESUR et les tués des BAAC, sur les variables pré-citées (force de l'ordre, type de réseau, type d'usager, tiers (oui/non) et âge (cf. tableau ci-après))

*Tableau 1 : Comparaison des tués des PV VOIESUR aux BAAC*

| variable<br>Test (p-value) | | VOIESUR<br>effectif | VOIESUR<br>% colonne | BAAC<br>%colonne |
|---|---|---|---|---|
| Force de l'ordre | CRS | 97 | 2,5 | 2,7 |
| p=0.37 | Gendarmerie | 2987 | 76,2 | 75,3 |
| | Police | 837 | 21,3 | 22,1 |
| | Total | 3921 | 100,0 | 100,0 |
| Réseau | Autoroute | 267 | 6,8 | 7,4 |
| p=0.07 | RN-RD | 2848 | 72,6 | 73,6 |
| | Voie Com. | 734 | 18,7 | 17,3 |
| | Autre | 72 | 1,8 | 1,7 |
| | Total | 3921 | 100,0 | 100,0 |
| Usager * tiers | 2RM avec tiers | 615 | 15,7 | 16,4 |
| p=0.61 | 2RM sans tiers | 321 | 8,2 | 8,3 |
| | vélo avec tiers | 124 | 3,2 | 3,1 |
| | Vélo sans tiers | 20 | 0,5 | 0,4 |
| | VL avec tiers | 1063 | 27,1 | 27,8 |
| | VL sans tiers | 1027 | 26,2 | 24,9 |
| | Piéton avec tiers | 505 | 12,9 | 13,1 |
| | Autre avec tiers | 122 | 3,1 | 2,9 |
| | Autre sans tiers | 124 | 3,2 | 3,1 |
| | Total | 3921 | 100,0 | 100,0 |
| Age | 0-13 ans | 90 | 2,3 | 2,7 |
| p=0.48 | 14-17 ans | 160 | 4,1 | 4,2 |
| | 18-24 ans | 775 | 20,0 | 20,5 |
| | 25-34 ans | 681 | 17,6 | 18,0 |
| | 35-69 ans | 1535 | 39,6 | 38,8 |
| | 70 ans et + | 639 | 16,5 | 15,8 |
| | Total (41 val mqt) | 3880 | 100,0 | 100,0 |

Les tests de comparaison de la proportion observée sur VOIESUR à celle des BAAC donnent des p-values supérieures au seuil de 5% ; on ne rejette donc pas l'égalité des répartitions.

De même, pour les blessés dans un accident mortel, il n'a pas été trouvé de différence significative entre les blessés dans les PV VOIESUR, et les blessés des BAAC, sur les mêmes variables (cf. tableau ci-après).

*Tableau 2 : comparaison des blessés dans les accidents mortels entre PV VOIESUR et BAAC*

| Variable<br>Test (p-value) | | VOIESUR<br>effectif | VOIESUR<br>% colonne | BAAC<br>% colonne |
|---|---|---|---|---|
| Force de l'ordre<br>p=0.33 | CRS | 58 | 2,5 | 2,7 |
| | Gendarmerie | 1907 | 81,0 | 79,8 |
| | Police | 389 | 16,5 | 17,5 |
| | Total | 2354 | 100,0 | 100,0 |
| Réseau<br>p=0.52 | Autoroute | 200 | 8,5 | 9,3 |
| | RN-RD | 1893 | 80,4 | 79,9 |
| | Voie Com. | 238 | 10,1 | 10,1 |
| | Autre | 23 | 1,0 | 0,8 |
| | Total | 2354 | 100,0 | 100,0 |
| Usager * tiers<br>p=0.40 | 2RM avec tiers | 92 | 3,9 | 4,5 |
| | 2RM sans tiers | 25 | 1,1 | 1,1 |
| | vélo avec tiers | 7 | 0,3 | 0,2 |
| | VL avec tiers | 1417 | 60,2 | 58,6 |
| | VL sans tiers | 461 | 19,6 | 19,8 |
| | Piéton avec tiers | 41 | 1,7 | 2,3 |
| | Autre avec tiers | 273 | 11,6 | 11,7 |
| | Autre sans tiers | 38 | 1,6 | 1,7 |
| | Total | 2354 | 100,0 | 100,0 |
| Age<br>p=0.58 | 0-13 ans | 154 | 6,7 | 7,4 |
| | 14-17 ans | 132 | 5,8 | 5,4 |
| | 18-24 ans | 551 | 24,1 | 25,0 |
| | 25-34 ans | 408 | 17,9 | 17,7 |
| | 35-69 ans | 862 | 37,7 | 37,1 |
| | 70 ans et + | 178 | 7,8 | 7,4 |
| | Total (69 val mqt) | 2285 | 100,0 | 100,0 |

Il n'y a donc pas lieu de redresser les tués et les blessés des PV mortels.

### 2.1.2 Comparaison PV VOIESUR – BAAC pour les accidents corporels, en termes de blessés

Pour les blessés dans les accidents corporels, il y a des différences significatives de répartition, en termes de force de l'ordre, type de réseau, type d'usager, accident avec ou sans tiers, et d'âge. Celles-ci sont détaillées dans le tableau ci-après.

Il est à noter que nous avons considéré les PV corporels au 1/20$^{ème}$ sur la France entière, i.e. avec Rhône au 1/20ème également (et non la totalité des PV du Rhône).

*Tableau 3: comparaison des blessés dans les accidents corporels entre PV 1/20ème VOIESUR et BAAC*

| Variable<br>Test (p-value) | | VOIESUR<br>effectif | VOIESUR<br>% colonne | BAAC<br>%colonne |
|---|---|---|---|---|
| Force de l'ordre | CRS | 275 | 7,2 | 6,5 |
| p<0.001 | Gendarmerie | 1236 | 32,3 | 25,6 |
| | Police | 2321 | 60,6 | 67,8 |
| | Total | 3832 | 100,0 | 100,0 |
| | | | | |
| Réseau | Autoroute | 339 | 8,9 | 7,6 |
| p<0.001 | RN-RD | 1641 | 42,8 | 38,6 |
| | Voie com. | 1761 | 46,0 | 51,6 |
| | Autre | 91 | 2,4 | 2,3 |
| | Total | 3832 | 3832 | 100,0 |
| | | | | |
| Usagers * tiers | 2RM avec tiers | 930 | 24,3 | 26,5 |
| p<0.001 | 2RM sans tiers | 209 | 5,5 | 6,1 |
| | vélo avec tiers | 182 | 4,8 | 5,0 |
| | Vélo sans tiers | 10 | 0,3 | 0,5 |
| | VL avec tiers | 1344 | 35,1 | 31,4 |
| | VL sans tiers | 442 | 11,5 | 11,1 |
| | Piéton avec tiers | 525 | 13,7 | 15,01 |
| | Autre avec tiers | 160 | 4,2 | 3,15 |
| | Autre sans tiers | 30 | 0,8 | 1,33 |
| | Total | 3832,0 | 100,0 | 100,01 |
| | | | | |
| Âge | 0-13 ans | 219 | 6,0 | 6,5 |
| p<0.02 | 14-17 ans | 255 | 7,0 | 7,8 |
| | 18-24 ans | 810 | 22,1 | 20,6 |
| | 25-34 ans | 697 | 19,0 | 20,4 |
| | 35-69 ans | 1451 | 39,6 | 38,4 |
| | 70 ans et + | 237 | 6,5 | 6,4 |
| | Total (163 val. mqt) | 3669 | 100,0 | 100,0 |

Les différences mises en évidence indiquent qu'il y a lieu de redresser.

## 2.2 Redressement des blessés des PV vers les BAAC

Redresser un échantillon signifie le rendre représentatif de la population étudiée.
Nous ne redressons pas sur toutes les variables où nous constatons une différence. En effet, redresser sur de nombreuses variables conduirait à perdre en précision. Nous choisissons de redresser sur type de force de l'ordre (CRS / gendarmerie / Police), type d'usager en cinq catégories (piétons / cyclistes / usagers de 2RM / automobilistes / autres) et tiers (oui/non). Nous pensons que redresser sur le type de force de l'ordre redressera en même temps sur le réseau, car cela est lié. De même, nous pensons que redresser sur type d'usager et tiers redressera en même temps sur l'âge.

Nous utilisons une méthode de redressement, par post-stratification (cf. annexe) (Deville and Sarndal 1992). Il s'agit de pondérer les observations de VOIESUR afin que la nouvelle répartition pondérée corresponde à la répartition connue sur les BAAC nationaux.

Les coefficients de redressement, se définissent donc, pour chaque combinaison des variables de redressement, par le ratio entre la proportion observée dans les BAAC et la proportion observée dans VOIESUR. Le tableau ci-après donne les répartitions des blessés selon les variables de redressement (type de force de l'ordre, type d'usager et tiers (oui/non)), dans les PV corporels au 1/20ème de VOIESUR, dans les BAAC, et les coefficients de redressement.

*Tableau 4 : distribution des blessés dans accidents corporels, VOIESUR PV au 1/20ème, et BAAC, selon les variables de redressement ; et coefficients de redressement*

| Force de l'ordre | Type d'usager * tiers | VOIESUR (n=3832) % colonne | BAAC (n=79000) % colonne | Coefficients de redressement (ratio) |
|---|---|---|---|---|
| CRS | 2RM avec tiers | 0,97 | 1,27 | 1,32 |
| CRS | 2RM sans tiers | 0,26 | 0,39 | 1,50 |
| CRS | vélo avec tiers | 0,03 | 0,01 | 0,19 |
| CRS | VL avec tiers | 3,78 | 3,11 | 0,82 |
| CRS | VL sans tiers | 1,46 | 1,20 | 0,82 |
| CRS | Autre avec tiers | 0,52 | 0,37 | 0,71 |
| CRS | Autre sans tiers | 0,16 | 0,17 | 1,07 |
| Gendarmerie | 2RM avec tiers | 5,45 | 4,74 | 0,87 |
| Gendarmerie | 2RM sans tiers | 1,80 | 1,90 | 1,05 |
| Gendarmerie | vélo avec tiers | 0,97 | 0,94 | 0,97 |
| Gendarmerie | Vélo sans tiers | 0,05 | 0,07 | 1,29 |
| Gendarmerie | VL avec tiers | 13,80 | 9,27 | 0,67 |
| Gendarmerie | VL sans tiers | 6,52 | 5,25 | 0,80 |
| Gendarmerie | Piéton avec tiers | 1,88 | 1,65 | 0,88 |
| Gendarmerie | Autre avec tiers | 1,33 | 1,12 | 0,84 |
| Gendarmerie | Autre sans tiers | 0,44 | 0,71 | 1,61 |
| Police | 2RM avec tiers | 17,85 | 20,50 | 1,15 |
| Police | 2RM sans tiers | 3,39 | 3,81 | 1,12 |
| Police | vélo avec tiers | 3,76 | 4,02 | 1,07 |
| Police | Vélo sans tiers | 0,21 | 0,40 | 1,92 |
| Police | VL avec tiers | 17,48 | 19,02 | 1,09 |
| Police | VL sans tiers | 3,55 | 4,62 | 1,30 |
| Police | Piéton avec tiers | 11,82 | 13,33 | 1,13 |
| Police | Autre avec tiers | 2,32 | 1,67 | 0,72 |
| Police | Autre sans tiers | 0,18 | 0,45 | 2,45 |
| | | 100,00 | 100,00 | |

Nous avons vérifié que les données des blessés VOIESUR une fois redressées avaient des répartitions similaires à celles des BAAC pour les variables type de réseau et âge, ne servant pas au redressement.

Pour l'âge il est à noter que le test global rejette l'hypothèse de répartitions égales, mais pour les classes d'âge « 0-13 ans » et « 70 ans et + » servant à définir des sous-populations pour les analyses de la tâche 4, l'intervalle de confiance de la proportion de blessés de VOIESUR englobe la proportion dans les BAAC (cf. tableau ci-après). Afin de ne pas multiplier les variables de redressement, nous considérons que cela est satisfaisant.

*Tableau 5: comparaison des blessés dans les accidents corporels entre les PV au 1/20ème, pondérés et redressés de VOIESUR et les BAAC*

| | | VOIESUR | | | BAAC |
|---|---|---|---|---|---|
| | | % colonne | IC à 95% | | % colonne |
| réseau | Autoroute | 7,83 | 7,00 | 8,65 | 7,56 |
| p=0.34 | RN-RD | 39,59 | 38,03 | 41,16 | 38,62 |
| | Voie com. | 50,19 | 48,57 | 51,80 | 51,58 |
| | Autre | 2,39 | 1,90 | 2,89 | 2,25 |
| | Total | 100,00 | | | 100,01 |
| Age | 0-13 ans | 5,94 | 5,16 | 6,73 | 6,48 |
| p=0.006 | 14-17 ans | 7,07 | 6,22 | 7,92 | 7,8 |
| | 18-24 ans | 22,52 | 21,12 | 23,91 | 20,55 |
| | 25-34 ans | 18,98 | 17,69 | 20,28 | 20,36 |
| | 35-69 ans | 39,38 | 37,76 | 40,99 | 38,43 |
| | 70 ans et + | 6,12 | 5,34 | 6,89 | 6,38 |
| | Total | 100,00 | | | 100,01 |

Contrainte sur les effectifs totaux :

Il s'est avéré qu'en appliquant les poids de sondage (20 ou 1) et les coefficients de redressement sur les données VOIESUR, nous obtenions un effectif moindre de blessés que l'effectif des BAAC (environ 79000 au lieu de 81000).
Etant donné que les PV redressés sont ensuite extrapolés, avec des coefficients du sous-enregistrement et des biais, basés sur les BAAC (afin d'estimer le bilan national de l'insécurité routière) nous avons ajouté des contraintes sur les effectifs redressés, afin d'obtenir les effectifs BAAC.

Nous ne l'avons fait que sur les blessés, qu'ils soient dans des accidents mortels ou corporels (pour les indemnes, ou ceux de gravité inconnue : ils sont traités comme les blessés).
Nous ne l'avons pas fait sur les tués ; en effet, il nous parait important de laisser apparent qu'il y a un peu plus de tués selon VOIESUR que selon les BAAC (environ 40 pour 4000).

## 3 EXTRAPOLATION : VERS UN BILAN NATIONAL CORRIGE DU SOUS-ENREGISTREMENT ET DE SES BIAIS

Nous reprenons une méthodologie que nous avons développée et mise en place précédemment (Amoros 2007; Amoros, Martin et al. 2008; Amoros, Martin et al. 2008), sur les données d'accidentalité de 1996 à 2004. Ici, nous reprenons cette méthodologie, pour l'appliquer sur les données 2011 de VOIESUR. Cette méthodologie doit être adaptée à plusieurs éléments.

Premièrement, il faut prendre en compte la nouvelle définition de blessé grave/ blessé léger dans les BAAC depuis 2005 : ils étaient auparavant définis comme hospitalisé plus de 6 jours / moins de 6 jours, et ils sont dorénavant définis comme hospitalisé (plus de 24h) / hospitalisé moins de 24h ou non hospitalisé. Pour simplifier nous écrirons simplement : hospitalisé / non-hospitalisé.

Deuxièmement, nous faisons un autre choix pour le critère de gravité commun entre le Registre du Rhône et les BAAC. Jusqu'à présent nous privilégions un critère médical, ce qui impliquait que nous prédisions ce critère sur les données BAAC, et pour cela, nous passions de données individuelles à des données agrégées. Il n'est en effet pas réaliste de prédire la gravité lésionnelle au niveau de l'individu. Nous disposons dorénavant d'un même critère de gravité pour les BAAC et le Registre du Rhône, celui d'hospitalisé (oui/non). Cela évite une étape de prédiction sur les données BAAC et donc de l'incertitude. Le passage du critère d'hospitalisé (oui/non) à un critère purement lésionnel (en l'occurrence le MAIS) est reporté à une étape ultérieure et optionnelle à la procédure d'extrapolation elle-même.

Troisièmement, la procédure d'extrapolation a besoin d'être mise à jour sur les années récentes, car elle portait sur 1996-2004. En effet, sur la période 1996-2004, nous avions mis en évidence un effet année, c'est-à-dire que l'ampleur du sous-enregistrement évoluait dans le temps. Il nous faut donc vérifier si cette évolution est la même et le cas échéant évaluer cette nouvelle évolution.
Cette procédure nous permet d'estimer des coefficients correcteurs du sous-enregistrement des BAAC, et de leurs biais associés sur la période 2006-2012. Nous utiliserons les coefficients correcteurs estimés pour l'année 2011 pour le projet VOIESUR (tâche 4).

La procédure d'extrapolation recouvre deux approches : l'approche capture-recapture suivie d'une méthode de projection.

L'approche de capture-recapture est d'abord utilisée au niveau du département du Rhône, où les victimes de la route sont enregistrées par deux sources : d'une part par les forces de l'ordre (et sont informatisées sous forme de BAAC), et d'autre part par le registre des victimes d'accidents de la route dans le Rhône. Le chainage entre les 2 sources, c'est-dire l'identification des victimes communes aux BAAC et au registre, permet aussi de voir que certaines victimes ne sont enregistrées que par une seule source. Cette défaillance dans l'enregistrement de l'une ou de l'autre source, indique que des victimes peuvent échapper aux deux enregistrements.
La méthode de capture-recapture, peut alors, sous certaines conditions, estimer le nombre de personnes échappant aux deux recensements. On estime ainsi le nombre réel de victimes de la route dans le Rhône (cf figure ci-dessous : sur-ensemble orange).

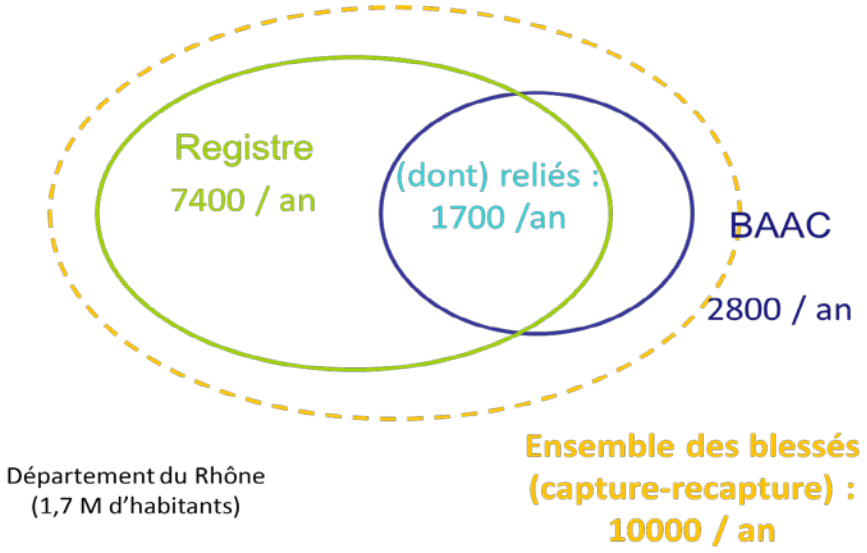


*Figure 2 : capture-recapture, département du Rhône – Effectifs annuels moyens 2006-2012*

Cette estimation de l'effectif du sur-ensemble se fait directement, par modélisation. Nous pouvons alors en déduire des coefficients de correction entre les effectifs des BAAC du Rhône et l'effectif du sur-ensemble total.

Intervient alors la méthode de projection : nous appliquons les coefficients de correction des BAAC, estimés sur le Rhône, aux BAAC nationaux (figure ci-dessous). Cela ne peut se faire que sous l'hypothèse d'homogénéité des pratiques d'enregistrement des blessés de la route par les forces de l'ordre sur la France entière. Autrement dit : le taux d'enregistrement des blessés de la route est environ le même pour les différents départements. Nous pensons que cette condition est relativement remplie, pour un type de force de l'ordre donnée (CRS, police, gendarmerie), chacune ayant une structure nationale.

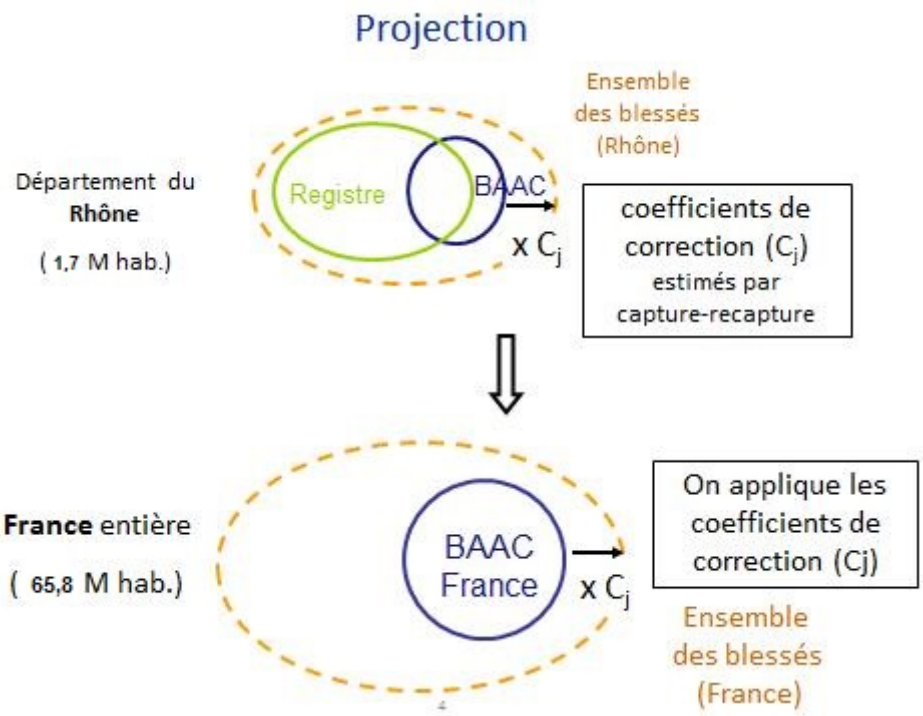


*Figure 3: projection, du Rhône vers la France métropolitaine (effectifs de population en 2014)*

## 3.1 Méthodologie de l'extrapolation en 4 étapes

Cette méthodologie de l'extrapolation se découpe en plusieurs étapes. La figure ci-après permet d'identifier dans le cadre rose ce qui modélisé au niveau des données rhodaniennes, et ce qui appliqué aux données nationales, en l'occurrence aux BAAC nationaux (cadre bleu) ou en particulier aux PV corporels au 1/20$^{ème}$ (cadre jaune représentant le projet VOIESUR). Dans le cadre rose, l'étape 1 est simplement un préambule à l'étape 2 de capture-recapture proprement dite, et l'étape 3 est la projection à l'échelle nationale. L'étape 4 est une étape optionnelle de prédiction de la gravité lésionnelle, qui permet d'obtenir des résultats en fonction du MAIS, critère médical, basé sur

l'Abbreviated Injury Scale, échelle traumatologique internationale. Cependant ces résultats ne sont que sous forme d'effectifs (la prédiction individuelle du MAIS n'est pas réaliste).

L'étape 1 consiste à améliorer le résultat du chaînage entre BAAC et Registre, rhodaniens, pour s'approcher de la condition n°2 de capture-recapture, c'est-à-dire l'identification parfaite des individus communs aux deux sources. Nous estimons le nombre de faux positifs et faux négatifs (définis en regard du chaînage) et nous corrigeons le résultat du chainage à l'aide de ces effectifs.

L'étape 2 est l'étape de capture-recapture elle-même, au niveau rhodanien : estimer le nombre total de blessés de la route, directement par le modèle, et en déduire les coefficients multiplicateurs de correction (du sous-enregistrement et de ses biais) entre les BAAC, et le sur-ensemble exhaustif obtenu.

L'étape 3 est l'étape de projection à l'échelle nationale : appliquer ces coefficients de correction (estimés au niveau rhodanien) aux BAAC nationaux, ou aux PV corporels au 1/20$^{ème}$ redressés, et permettre ainsi une estimation exhaustive et représentative du bilan national des blessés de la route.

L'étape 4 est une étape optionnelle qui permet de passer de données décrites avec le critère hospitalisé (oui/non) et d'autres caractéristiques des accidents et des blessés, à des données, agrégées cependant, décrites avec un critère médical de la gravité lésionnelle, le MAIS3+.

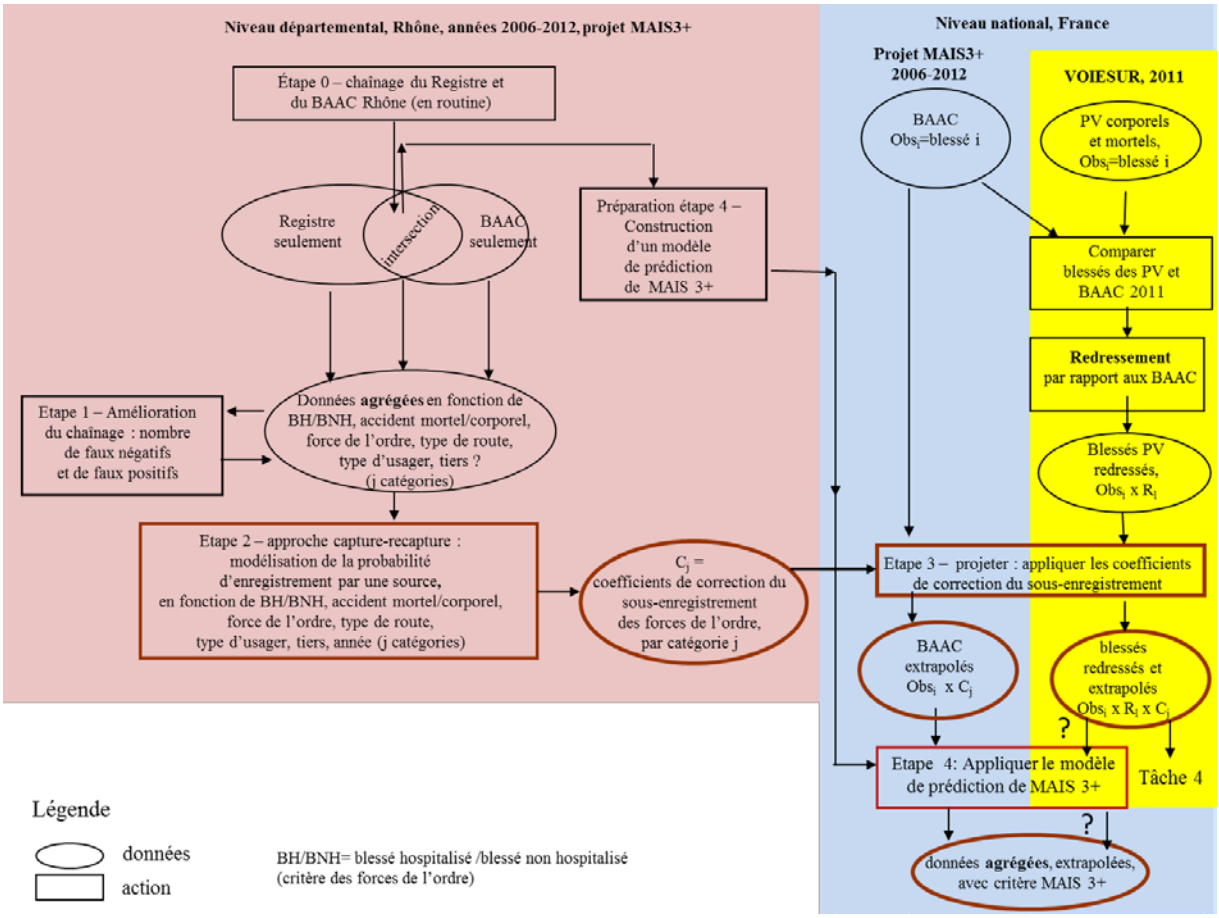


*Figure 4 : description des étapes de création des poids pour redresser et extrapoler*

## 3.2 Situation de l'enregistrement dans le Rhône

Globalement sur la période 2006-2012, les BAAC ont enregistré 2800 blessés par an, et le Registre 7400. Les blessés identifiés comme communs aux deux sources (« reliés ») sont au nombre de 1700 par an. L'ensemble total est donc de 8500 blessés/an. Cela correspond à 3 fois plus que les BAAC.

Les tableaux ci-après indiquent la situation de l'enregistrement des blessés de la route dans les BAAC et dans le Registre au niveau du Rhône, en les déclinant selon les variables qui jouent le plus sur la probabilité d'enregistrement.

Pour les blessés identifiés par le Registre seulement, la force de l'ordre « en charge » de la zone est déterminée par la commune et le type de réseau où a eu lieu l'accident.

Le tableau ci-après montre que la propension des forces de l'ordre à enregistrer les blessés n'est pas la même d'un type de force de l'ordre à l'autre. Ainsi, il suffirait de multiplier les données BAAC par 1,7 en zone CRS, 2,8 en zone Police mais par 5,3 en zone gendarmerie (pour obtenir l'ensemble BAAC U Registre).

*Tableau 6 : effectifs annuels moyens des blessés selon la source et le type de force de l'ordre, Rhône, 2006-2012*

| | BAAC | Registre | Reliés | Ensemble | Ratio (Ens/BAAC) |
|---|---|---|---|---|---|
| CRS | 524 | 673 | 305 | 892 | 1,7 |
| Gendarmerie | 549 | 2690 | 355 | 2884 | 5,3 |
| Police | 1712 | 4054 | 1046 | 4720 | 2,8 |
| Total | 2785 | 7417 | 1706 | 8496 | 3,1 |

Ces disparités entre types de forces de l'ordre, recouvrent aussi des disparités selon le classement hospitalisé (plus de 24h) (oui/non). Ainsi, comme le montre le tableau ci-après, pour les blessés hospitalisés, il n'y a pas tant d'écart d'enregistrement entre les 3 types de forces de l'ordre, mais pour les blessés non hospitalisés la propension à les enregistrer est très différente.

*Tableau 7 : effectifs annuels moyens des blessés selon la source, le type de force de l'ordre et la gravité des blessés, Rhône, 2006-2012*

| | BAAC | Registre | reliés | Ensemble | Ratio (E/B) |
|---|---|---|---|---|---|
| CRS, bl. hospitalisé | 104 | 94 | 71 | 128 | 1,2 |
| CRS, bl. non hosp. | 420 | 579 | 234 | 765 | 1,8 |
| Gend., bl. hospitalisé | 345 | 527 | 244 | 627 | 1,8 |
| Gend., bl non hosp. | 204 | 2163 | 110 | 2257 | 11,0 |
| Police, bl. hospistalisé | 419 | 517 | 301 | 635 | 1,5 |
| Police, bl. non hosp. | 1293 | 3538 | 745 | 4085 | 3,2 |
| Total | 2785 | 7417 | 1706 | 8496 | 3,1 |

La probabilité d'enregistrement dans les BAAC dépend aussi du type d'usager et de l'existence ou non d'un tiers dans l'accident. Le tableau ci-après donne les effectifs selon les BAAC, le Registre, et l'ensemble. Les accidents sans tiers sont bien moins enregistrés dans les BAAC que les accidents avec tiers, et l'écart dépend du type d'usager. On note le cas très particulier des cyclistes blessés dans un accident sans tiers : ils sont très rarement enregistrés dans les BAAC.

***Tableau 8** : effectifs annuels moyens des blessés selon la source, et selon le type d'usager et la présence d'un tiers, Rhône, 2006-2012*

| | BAAC | Registre | Reliés | Ensemble | Ratio (E/B) |
|---|---|---|---|---|---|
| 2RM, avec tiers | 582 | 918 | 390 | 1110 | 1,9 |
| 2RM, sans tiers | 100 | 1050 | 65 | 1085 | 10,9 |
| Vélo, avec tiers | 138 | 301 | 88 | 351 | 2,5 |
| Vélo, sans tiers | 6 | 855 | 4 | 857 | 136,3 |
| VL, avec tiers | 1072 | 2328 | 619 | 2780 | 2,6 |
| VL, sans tiers | 273 | 1012 | 177 | 1108 | 4,1 |
| Piéton, avec tiers | 449 | 705 | 277 | 877 | 2,0 |
| Autre, avec tiers | 117 | 114 | 61 | 170 | 1,4 |
| Autre, sans tiers | 48 | 134 | 24 | 158 | 3,3 |
| Total | 2785 | 7417 | 1706 | 8496 | 3,1 |

Enfin, la probabilité d'enregistrement des blessés dans les BAAC dépend du type de réseau, pour certaines forces de l'ordre. Cela ne semble pas jouer en zone CRS, mais ils ne s'occupent que des accidents sur autoroutes ou voies rapides. Cela joue peu en zone police, mais en revanche, en zone gendarmerie, les blessés sur voies communales sont bien moins enregistrés que les blessés sur les autres réseaux.

***Tableau 9 :** effectifs annuels moyens des blessés selon la source, et selon la force de l'ordre et le réseau, Rhône, 2006-2012*

| | BAAC | Registre | reliés | Ensemble | Ratio (E/B) |
|---|---|---|---|---|---|
| CRS AR | 334 | 432 | 196 | 571 | 1,7 |
| CRS RN-RD | 177 | 236 | 104 | 309 | 1,7 |
| Gend AR | 27 | 44 | 17 | 55 | 2,0 |
| Gend RN-RD | 383 | 1190 | 254 | 1318 | 3,4 |
| Gend VC | 123 | 1388 | 75 | 1436 | 11,7 |
| Gend autre | 16 | 68 | 9 | 75 | 4,6 |
| Police RN-RD | 117 | 212 | 72 | 257 | 2,2 |
| Police VC | 1553 | 3755 | 949 | 4359 | 2,8 |
| Police autre | 38 | 85 | 23 | 100 | 2,6 |
| Total | 2785 | 7417 | 1706 | 8496 | 3,1 |

## 3.3 Etape 1 : Amélioration du chaînage entre BAAC et Registre, Rhône

Avant de modéliser le sous-enregistrement par l'approche capture-recapture, étape 2, il est nécessaire de nous approcher au plus près d'une des conditions de cette méthode, à savoir, l'identification parfaite des sujets communs aux deux sources, ici, BAAC et Registre rhodaniens.
Cela étant difficilement atteignable, et difficilement mesurable, nous estimons de façon aussi fiable que possible, la répartition des blessés selon leur source d'enregistrement, en distinguant :
- les blessés communs aux BAAC et au Registre, rhodaniens, notés « intersection BAAC-Registre »,
- les blessés présents seulement dans les BAAC, notés « BAAC seulement »,
- les blessés présents seulement dans le Registre, notés « Registre seulement ».

Nous travaillons là-aussi en termes d'effectifs (de blessés), i.e. de données agrégées, et non sur les blessés eux-mêmes, i.e. les données individuelles.

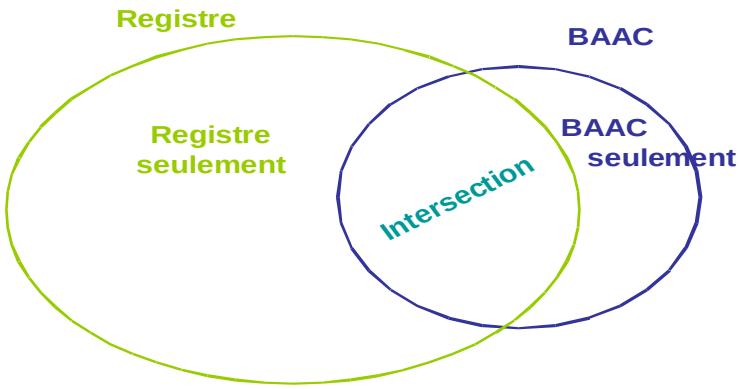


*Figure 5 : répartition des blessés recensés par le Registre et par les BAAC, département du Rhône*

Lors de la procédure de rapprochement ou chaînage (record-linkage), on compare les victimes des BAAC à celles du Registre (Rhône) en formant des paires constituées d'un enregistrement d'une victime des BAAC et d'un enregistrement d'une victime du Registre. Ces paires peuvent être classées de la façon suivante, dans le tableau ci-après.

*Tableau 10 : description d'un résultat de chaînage entre deux fichiers d'enregistrement (unité=une paire d'enregistrements)*

| décision : | situation réelle : enregistrements de : même individu (Matched) | individus différents (Unmatched) |
|---|---|---|
| rapprochement | vrais positifs | faux positifs |
| non-rapprochement | faux négatifs | vrais négatifs |

Positif/négatif réfère au résultat de la décision : positif signifie rapproché ou « relié » ou « chaîné », et négatif signifie non-rapproché. Faux et vrai qualifient la décision par rapport à la réalité. Dans la terminologie du record-linkage, l'ensemble des paires d'enregistrements correspondant (réellement) au même individu est noté M pour Matched, et l'ensemble des paires d'enregistrements correspondant (réellement) à des individus distincts est noté U pour Unmatched. Matched et Unmatched constituent ensemble le gold standard. Enfin, « relié » se traduit par « linked ».

Les faux positifs sont les paires d'enregistrements reliées à tort, car ne correspondant pas à la même victime. Les faux négatifs sont les paires non-reliées, à tort, c'est-à-dire deux enregistrements non reliés alors qu'ils correspondent à la même victime (Clark 2004).

Nous allons améliorer le résultat du chaînage par l'estimation du nombre de faux positifs et de faux négatifs.

### 3.3.1 Estimation du nombre de faux positifs

Un faux positif est une paire d'enregistrements de 2 victimes distinctes que l'on a reliée (à tort donc).

P(faux positifs)
=P(rapprocher alors qu'il s'agit d'enregistrements de 2 victimes distinctes)
=P(rapprocher / victimes distinctes)
=P(concordance sur toutes ou une partie des variables de chaînage / victimes distinctes)

La concordance sur toutes ou certaines variables est précisément définie. Cela correspond aux critères de décision de la procédure de rapprochement.

Les variables de chaînage sont les suivantes :

1) jour de l'accident
2) mois de l'accident
- (année de l'accident)
3) lieu de l'accident
4) mois de naissance de la victime
5) année de naissance de la victime
6) sexe de la victime
7) type d'usager de la victime

Remarque 1 : l'année de l'accident n'est pas considérée comme une variable de chaînage car elle est en fait utilisée comme variable de blocage, c'est-à-dire que la procédure de chaînage est conduite par blocs, qui sont définis par les années. Cela suppose l'absence d'erreur sur cette variable. Une précaution est quand même prise : pour les accidents ayant lieu en fin d'année (d'après les BAAC), les enregistrements du Registre des jours du début de l'année suivante sont explorés.

Remarque 2 : si lors de la comparaison des modalités BAAC et Registre d'une variable de chaînage, il y a une valeur manquante, nous considérons que le résultat de la comparaison est discordant.

Nous décrivons ce qui se passait sur la période 1996-2001 (car cela va nous servir pour la période 2006-2012) ; la décision de relier la paire se faisait si :
- les sept variables de chaînage concordent
- seulement six variables concordent
- seulement cinq variables concordent, mais les deux variables discordantes sont des variables considérées comme « mineures » : mois de naissance, sexe ou type d'usager de la victime.

Il y a eu un changement de stratégie de chaînage à partir de 2002. A partir de cette date, le chaînage entre les données des forces de l'ordre et le Registre se fait lors de la saisie des fiches de notification du Registre, avec l'idée, s'il y a chainage pour la victime en cours de saisie, de récupérer une partie des informations des forces de l'ordre afin de les incorporer à l'enregistrement en cours côté

Registre : notamment les informations sur les circonstances de l'accident (mais non les caractéristiques du blessé), jugées plus fiables. Cela conduit à une décision de rapprochement qui est plus restrictive. Concrètement : seulement une discordance sur une seule variable est tolérée (et cela ne doit pas porter sur le mois de l'accident).

En résumé, depuis 2002, on décide de rapprocher deux enregistrements s'il y a :
- concordance sur les sept variables de chaînage, ou
- concordance sur six variables seulement (et la discordance ne concerne pas le mois de l'accident)

Ces deux situations étant disjointes, la probabilité de l'ensemble est la somme des probabilités des deux situations :
P(rapprocher/ victimes distinctes)
$= P(7 \text{ var concordent / victimes distinctes}) + P(\text{exact. } 6 \text{ var concordent (sf mois) / victimes distinctes})$

Pour chacune des deux situations (écrites sur 2 lignes ci-dessous), en supposant l'indépendance des variables de chaînage, la probabilité de chacune s'écrit comme le produit des probabilités de concordance pour chaque variable de chaînage (entre les deux enregistrements), et multipliée, dans la 2ème situation, par la probabilité de discordance sur une seule variable, autre que le mois de l'accident :

$$= \left[ \prod_{i=1}^{i=7} P(\text{var}_i \text{ concorde/ victimes distinctes}) \right]$$

$$+ \left[ \sum_{j=1, j\neq 2}^{7} \left( \prod_{i=1, i\neq j}^{i=7} P(\text{var}_i \text{concorde/ victimes distinctes}) \times P(\text{var}_j \text{discorde/ victimes distinctes}) \right) \right]$$

La concordance sur une variable alors qu'il s'agit de victimes distinctes correspond à une concordance par simple hasard (par exemple pour la variable type d'usager, ce seraient deux automobilistes). Pour les variables mois et jour (de naissance ou d'accident) pour lesquelles les modalités sont uniformément réparties (à condition que les effectifs soit suffisants), les probabilités de concordance par hasard seront de 1/12 (i.e. 0.083) pour le mois et environ 1/30 ( i.e. 0.033) pour le jour (Brenner and Schmidtmann 1996).

Pour les autres variables, si l'on dispose du gold-standard (ensembles Matched et Unmatched) alors les proportions de concordance sont estimées par les proportions observées sur l'ensemble (Unmatched) des paires d'enregistrements de victimes (réellement) distinctes.
Comme nous ne le connaissons pas, nous choisissons d'approximer par l'ensemble des paires non-reliées, voire par l'ensemble de toutes les paires possibles, car la différence entre ces deux sous-ensembles est négligeable. En effet, les tailles respectives sont (avec « # » signifiant effectif, et « ∩ » signifiant intersection des 2 ensembles, i.e. blessés communs aux deux sources):

# paires non-reliées=#BAAC * #Registre - #(BAAC∩Registre) et
# toutes les paires possibles= #BAAC * #Registre,
avec #(BAAC∩Registre) négligeable devant #BAAC * #Registre.

Nous choisissons d'approximer par l'ensemble des paires non-reliées, qui comprend quelques dizaines de millions de paires (cf. tableau ci-après).

*Tableau 11 : proportions de concordance sur les variables de chaînage, observées sur l'ensemble des paires non reliées, registre et BAAC, Rhône, 1996-2004 et 2006-2012*

| variables chaînage | 1996 (n=45,4M) | 1997 (n=43,6M) | 1998 (n=50,2M) | 1999 (n=51,9M) | 2000 (n=44,1) | 2001 (n=43,9M) | 2002 (n=33,6M) | 2003 (n=25,4M) | 2004 (n=23,8M) |
|---|---|---|---|---|---|---|---|---|---|
| jour accident | 3,3% | 3,2% | 3,3% | 3,3% | 3,3% | 3,2% | 3,3% | 3,3% | 3,3% |
| mois accident | 8,4% | 8,5% | 8,4% | 8,4% | 8,4% | 8,4% | 8,4% | 8,5% | 8,5% |
| *année accident* | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |
| lieu de l'accident (commune) | 9,0% | 8,1% | 10,6% | 10,3% | 8,1% | 8,1% | 6,8% | 6,4% | 5,6% |
| mois de naissance | 8,3% | 8,3% | 8,3% | 8,3% | 8,3% | 8,3% | 8,2% | 8,2% | 8,3% |
| année de naissance | 1,9% | 1,9% | 1,9% | 1,9% | 2,0% | 2,0% | 1,9% | 1,9% | 1,8% |
| sexe | 52,1% | 52,5% | 52,0% | 52,3% | 52,5% | 52,4% | 52,5% | 53,3% | 52,6% |
| type d'usager | 38,4% | 38,5% | 40,2% | 37,8% | 38,3% | 39,2% | 38,0% | 33,3% | 33,0% |

| variables chaînage | 2006 (n=22,5M) | 2007 (n=21,8M) | 2008 (n=18,6M) | 2009 (n=21,5M) | 2010 (n=21,6) | 2011 (n=18,5M) | 2012 (n=20,2M) |
|---|---|---|---|---|---|---|---|
| jour accident | 3,30% | 3,30% | 3,20% | 3,30% | 3,20% | 3,30% | 3,20% |
| mois accident | 8,50% | 8,50% | 8,50% | 8,50% | 8,60% | 8,50% | 8,40% |
| *année accident* | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% | 100,00% |
| lieu de l'accident (commune) | 8,70% | 10,50% | 9,80% | 9,30% | 9,10% | 9,80% | 9,10% |
| mois de naissance | 8,30% | 8,40% | 8,40% | 8,40% | 8,40% | 8,40% | 8,30% |
| année de naissance | 1,90% | 2,10% | 1,90% | 2,00% | 1,90% | 1,80% | 1,80% |
| sexe | 51,70% | 52,90% | 52,30% | 52,80% | 52,10% | 52,30% | 52,30% |
| type d'usager | 32,30% | 32,20% | 30,10% | 30,30% | 30,10% | 29,60% | 30,70% |

Nous vérifions que la proportion observée de concordance sur le jour de l'accident (3,2%) est effectivement proche de 1/30 (qui vaut 3,33%), et celles observées sur le mois d'accident (8,4%), et le mois de naissance (8,3%) sont effectivement proches de 1/12, valant 8,33%.

Les proportions de concordance (par hasard) sont liées à la répartition des modalités. La baisse de la proportion de concordance sur le type d'usager peut s'expliquer par une baisse de la proportion d'automobilistes, qui est la catégorie la plus fréquente ; la variable type d'usager devient ainsi plus discriminante.

Concernant le lieu de l'accident, nous minorons la proportion obtenue car celle-ci est la concordance par hasard sur la commune, alors que c'est la concordance sur commune ET route ou rue qui est utilisée dans le chaînage ; celle-ci est bien plus faible. Un taux de concordance par hasard est arbitrairement fixé à 1%.

**Calcul du nombre de faux positifs :**

Nous notons :

FP = faux positifs
FN = faux négatifs
VN =vrais négatifs
VP = vrais positifs
#= effectif

alors :

#FP = P(FP) × #VN
#FP = P(FP) × (#BAAC × #Registre – #VP)
#FP = P(FP) × #BAAC × #Registre – P(FP) × #VP

le dernier terme est négligeable (de l'ordre de 10.E-3 en regard d'un effectif)
d'où : #FP = P(FP) × #BAAC × #Registre

Les résultats sur 1996-2004 et 2006-2012 sont donnés dans le tableau ci-après.

*Tableau 12 : proportions de concordance sur U (=Unmatched), de faux positifs et effectifs de faux positifs ; résultats du chaînage Registre et BAAC, Rhône, par année, sur 1996-2004 et 2006-2012*

| critère | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|---|---|
| P(FP) | 2,0 E-6 | 2,1 E-6 | 2,1 E-6 | 2,0 E-6 | 2,1 E-6 | 2,1 E-6 | 1,7 E-6 | 1,5 E-6 | 1,4 E-6 |
| # FP | 92 | 90 | 107 | 105 | 94 | 92 | 56 | 38 | 34 |

| critère | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|
| P(FP) | 1,5 E-6 | 1,5 E-6 | 1,4 E-6 | 1,4 E-6 | 1,4 E-6 | 1,3 E-6 | 1,3 E-6 |
| # FP | 32,8 | 33,7 | 25,3 | 30,3 | 29,1 | 24,2 | 26,6 |

On constate que la probabilité de faux positifs est légèrement plus faible à partir de 2002, date à laquelle la stratégie de chaînage devient plus restrictive : on retrouve dans les valeurs estimées l'idée logique, que si le chainage est plus restrictif, la proportion de faux positifs diminue. A l'inverse, la proportion de faux négatifs devrait augmenter.

### 3.3.2 Estimation du nombre de faux négatifs

Un faux négatif est une paire d'enregistrements de la même victime que l'on n'a pas rapprochée (à tort donc).

P(faux négatif)

=P(ne pas rapprocher alors qu'il s'agit de 2 enregistrements d'une même victime)
=P(ne pas rapprocher / même victime)
=P(discordance sur toutes ou une partie des variables de chaînage / même victime)
=1-P(concordance sur toutes ou une partie des variables de chaînage / même victime)

« Concordance sur toutes ou une partie des variables de chaînage » est précisément défini. Cela correspond aux critères de décision de la procédure de rapprochement. Nous les rappelons ci-après.

Sur la période 1996-2001, la décision de relier la paire était prise si :
- les sept variables de chaînage concordent
- seulement six variables concordent
- seulement cinq variables concordent, mais les deux variables discordantes sont des variables considérées comme « mineures » : mois de naissance, sexe ou type d'usager de la victime.

Depuis 2002, on décide un lien entre deux enregistrements s'il y a :
- concordance sur les 7 variables, ou
- concordance sur 6 variables seulement (et pas de discordance sur le mois de l'accident)

Ces deux situations étant disjointes, la probabilité de l'ensemble est la somme des probabilités des deux situations. En supposant l'indépendance des variables de chaînage, la probabilité de chacune des 2 situations s'écrit comme le produit des probabilités de concordance pour chaque variable de chaînage (entre les deux enregistrements), multipliée, dans la 2ème situation par la probabilité de discordance sur une seule variable, autre que le mois de l'accident :

P(concordance sur toutes ou une partie des variables de chaînage / même victime)

$$= P(7 \text{ variables concordent / même victime}) + P(6 \text{ variables concordent / même victime})$$

$$= \left[ \prod_{i=1}^{i=7} P(\text{var}_i \text{concorde / même victime}) \right] + \left[ \sum_{j=1, j\neq 2}^{7} \left( \prod_{i=1, i\neq j}^{i=7} P(\text{var}_i \text{concorde / même victime}) \times P(\text{var}_j \text{discorde / même victime}) \right) \right]$$

D'après Brenner et Schmidtmann, si on dispose du « gold standard », i.e. si on connait le sous-ensemble M des (vraies) paires correspondant à une même victime, on remplace ces probabilités par les proportions observées sur l'ensemble M.
Comme nous ne disposons pas du gold standard, nous pouvons utiliser le sous-ensemble des paires rapprochées comme une approximation grossière du sous-ensemble des paires des mêmes victimes. Mais cela donnera plutôt une sur-estimation des probabilités de concordance. Nous décidons donc de diminuer ces proportions.
Un autre élément conduit aussi à diminuer ces proportions : chacune de ces probabilités est égale à 1-taux d'erreur (Howe 1998), et cela ne parait pas réaliste d'avoir des taux d'erreur aussi faibles que 2 ou 3%. Dans la littérature, selon une étude citée par Howe (Howe 1998), le taux d'erreur sur l'année de naissance était de 13%.

***Tableau 13 : proportions de concordance sur quelques variables de chaînage, cité par Brenner, d'après Newcombe (Newcombe 1988),***

| variables de chaînage | concordance parmi les paires de même individu |
|---|---|
| nom | 96,5% |
| prénom | 79,0% |
| initiale du 2ème prénom | 88,8% |
| jour de naissance | 85,1% |
| mois de naissance | 93,3% |
| année de naissance | 77,3% |
| région de naissance | 98,1% |

Le tableau ci-après donne les proportions de concordance observées sur l'ensemble des paires reliées, par année.

*Tableau 14 : proportions de concordance sur les variables de chaînage, observées sur les paires reliées, Registre et BAAC, Rhône, 1996-2004*

| variables de chaînage | 1996 n=2919 | 1997 n=2616 | 1998 n=3169 | 1999 n=3040 | 2000 n=2761 | 2001 n=2809 | 2002 n=2013 | 2003 n= 1882 | 2004 n=1780 |
|---|---|---|---|---|---|---|---|---|---|
| jour acc. | 97,7% | 97,9% | 98,8% | 98,3% | 98,5% | 97,3% | 99,5% | 99,7% | 99,6% |
| mois acc | 99,9% | 99,8% | 99,9% | 99,8% | 99,8% | 99,9% | 100,0% | 100,0% | 100,0% |
| *année acc* | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% | 100,0% |
| lieu acc commune | 85,4% | 79,7% | 82,4% | 82,1% | 80,6% | 82,0% | 98,4% | 98,9% | 98,8% |
| mois de naissance | 88,5% | 95,9% | 94,5% | 95,5% | 95,8% | 95,8% | 97,2% | 97,3% | 98,4% |
| année de naissance | 92,3% | 97,0% | 95,8% | 96,2% | 97,0% | 96,5% | 97,8% | 98,4% | 98,1% |
| sexe | 96,9% | 96,3% | 96,9% | 95,9% | 96,0% | 97,0% | 98,6% | 98,1% | 98,7% |
| type d'usager | 95,9% | 94,3% | 95,6% | 95,7% | 95,1% | 96,1% | 98,0% | 95,7% | 98,7% |

Les proportions sont élevées, et encore plus à partir de 2002 (surtout pour la variable commune), où la décision de relier est plus restrictive. Pour que les proportions de ces années-là ne soient pas dépendantes du changement dans la stratégie de chaînage, nous utilisons les proportions de l'ensemble de la période 1996-2001 (ce n'était pas visible sur les paires non-reliées car le changement est dilué sur les quelques millions de paires non reliées). Ces proportions restent des proportions estimées sur les paires reliées, qui sur-estiment donc la probabilité de concordance sur les vrais positifs. Pour avoir des probabilités proches de celles de la littérature, nous minorons de 5% les proportions observées. Le tableau ci-après donne ces nouvelles valeurs.

*Tableau 15 : proportions de concordance sur les variables de chaînage, observées sur les paires reliées et minorées de 5%, Registre et BAAC, Rhône, 1996-2001*

| variables de chaînage | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | **moyenne** |
|---|---|---|---|---|---|---|---|
| jour accident | 93% | 93% | 94% | 93% | 94% | 92% | **93%** |
| mois acc | 95% | 95% | 95% | 95% | 95% | 95% | **95%** |
| *année acc* | 100% | 100% | 100% | 100% | 100% | 100% | **100%** |
| lieu accident (commune) | 81% | 76% | 78% | 78% | 77% | 78% | **78%** |
| mois de naissance | 84% | 91% | 90% | 91% | 91% | 91% | **90%** |
| année de naissance | 88% | 92% | 91% | 91% | 92% | 92% | **91%** |
| sexe | 92% | 91% | 92% | 91% | 91% | 92% | **92%** |
| type usager | 91% | 90% | 91% | 91% | 90% | 91% | **91%** |

**Calcul du nombre de faux négatifs :**

La probabilité de faux négatifs est définie sur l'ensemble des vrais positifs.

Avec les notations précédentes, on a :

$\#FN = P(FN) \times \#VP$

$\#FN = P(FN) \times (\#Pos - \#FP + \#FN)$

$(1 - P(FN)) \times \#FN = P(FN) \times (\#Pos - \#FP)$

$\#FN = P(FN) \times (\#Pos - \#FP) / (1 - P(FN))$

Le tableau ci-après donne les résultats sur 1996-2004 et 2006-2012.

*Tableau 16 : proportions de concordance sur M (=Matched), de faux négatifs, d'effectifs de faux négatifs ; résultats du chaînage Registre et BAAC, Rhône, par année, 1996-2004 et 2006-2012*

| critère | 1996 | 1997 | 1998 | 1999 | 2000 | 2001 | 2002 | 2003 | 2004 |
|---|---|---|---|---|---|---|---|---|---|
| P(concordance / ens. M) | 84,9% | 86,1% | 87,0% | 86,9% | 86,7% | 87,1% | 82,4% | 82,4% | 82,4% |
| P(FN) | 15,1% | 13,9% | 13,0% | 13,1% | 13,3% | 12,9% | 17,6% | 17,6% | 17,6% |
| # FN | 505 | 408 | 457 | 445 | 410 | 403 | 417 | 393 | 369 |

| critère | 2006 | 2007 | 2008 | 2009 | 2010 | 2011 | 2012 |
|---|---|---|---|---|---|---|---|
| P(concordance / ens. M) | 82,40% | 82,40% | 82,40% | 82,40% | 82,40% | 82,40% | 82,40% |
| P(FN) | 17,60% | 17,60% | 17,60% | 17,60% | 17,60% | 17,60% | 17,60% |
| # FN | 375 | 361 | 332 | 346 | 379 | 354 | 358 |

On note que les probabilités de faux négatifs sont plus élevées à partir de 2002, ce qui est attendu pour une stratégie de chainage plus restrictive.

### 3.3.3 Redéfinition des effectifs résultant du chaînage

Les effectifs estimés de faux négatifs et de faux positifs permettent d'améliorer le résultat du chaînage. Plus précisément il permet de redéfinir les tailles de chaque sous-ensemble : « BAAC seulement », « Registre seulement » et « intersection BAAC-Registre » (cf. figure ci-après).

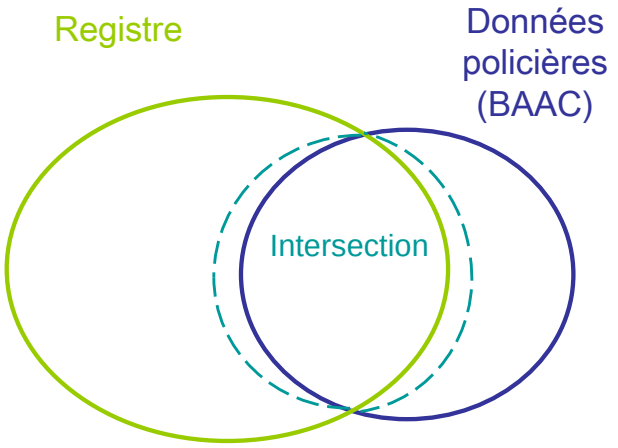


Figure 6 : redéfinition de la taille de l'intersection BAAC-Registre, Rhône

L'intersection BAAC-Registre est diminuée par l'effectif des faux positifs et augmentée de l'effectif des faux négatifs. Les sous-ensembles « BAAC seulement », « Registre seulement » sont chacun augmentés de l'effectif des faux positifs et diminués de l'effectif des faux négatifs.
Cela s'écrit (avec l'astérisque signifiant nouvel effectif) :

$$n^{*}_{\text{BAAC seul.}} = n_{\text{BAAC seul.}} - n_{\text{faux neg}} + n_{\text{faux pos}}$$

$$n^{*}_{\text{Reg. seul.}} = n_{\text{Reg. seul.}} - n_{\text{faux neg}} + n_{\text{faux pos}}$$

$$n^{*}_{\text{intersection}} = n_{\text{intersection}} + n_{\text{faux neg}} - n_{\text{faux pos}}$$

Un faux négatif est une paire constituée d'un individu de « BAAC seulement » et d'un individu de « Registre seulement ». Ainsi l'effectif de faux négatifs est égal à l'effectif d'individus de « BAAC seulement » non reliés à tort, et aussi égal à l'effectif d'individus de « Registre seulement » non reliés à tort. De même, l'effectif de faux positifs est un nombre de paires, mais aussi égal au nombre de blessés du Registre reliés à tort, et égal au nombre de blessés des BAAC reliés à tort.

Nous rappelons que nous travaillons sur des données agrégées, et que, à ce stade, nous n'avons besoin que de la variable d'appartenance à l'un des trois sous-ensembles et des variables qui seront utilisées dans le modèle de capture-recapture. Sur la période 2006-2012 : il s'agit du critère hospitalisé (oui/non), la gravité de l'accident (mortel/corporel), le type de force de l'ordre, le type d'usager, le tiers (oui/non), le type de réseau, et l'année (principaux facteurs de biais de sélection).

Nous notons les sous-ensembles exclusifs de la façon suivante :
- k=1 « BAAC seulement »
- k=2 « Registre seulement »
- k=3 « intersection BAAC-Registre »

Nous notons j les strates définies par la combinaison des principaux facteurs de biais de sélection. Nous notons $p_{kj}$ la répartition des strates j dans le sous-ensemble k, c'est-à-dire :

$$p_{kj} = n_{kj} / n_k$$

Dans chacun des trois sous-ensembles exclusifs, nous avons besoin de la nouvelle répartition des facteurs de biais de sélection. Pour un sous-ensemble donné, celle-ci est supposée être la même qu'avant la correction par les effectifs de faux négatifs et de faux positifs ; en d'autres termes :

$$p^*_{kj} = p_{kj}$$

Cette hypothèse vient du fait que les faux négatifs et les faux positifs ne sont ainsi que par manque d'efficacité de la procédure de chaînage. Par exemple : les blessés qui sont réellement dans l'intersection BAAC-Registre mais n'avaient pas été identifiés comme tels par la procédure de chaînage (pour cause d'erreurs dans les variables de chaînage) ressemblent aux blessés de l'intersection (en termes de caractéristiques de l'accident et du blessé) et non pas aux blessés de « BAAC seulement » ou de « Registre seulement ».

Dans chaque sous-ensemble, il suffit d'appliquer, au nouvel effectif, la répartition (croisée de la gravité du blessé, de l'accident, du type force de l'ordre, de type d'usager, tiers, réseau et année) observée avant la réaffectation de ceux qui appartiennent à ce groupe mais n'avaient pas été identifiés comme tels, c'est-à-dire :

$$n^*_{kj} = p_{kj} \times n^*_k$$

Le tableau ci-après donne un exemple des effectifs corrigés des faux négatifs et faux positifs du chaînage.

*Tableau 17 : effectifs moyens annuels de blessés selon la source, et corrigées des nombres de faux positifs et faux négatifs du chaînage, 2006-2012*

| *Type d'usager et tiers (oui/non)* | BAAC | Registre | Intersection (reliés) | BAAC U Registre (ensemble) | corrigé des FN et FP |
|---|---|---|---|---|---|
| 2RM avec Tiers | 582 | 918 | 390 | 1110 | 1098 |
| 2RM sans Tiers | 100 | 1050 | 65 | 1085 | 1031 |
| vélo Avec tiers | 138 | 301 | 88 | 351 | 341 |
| vélo Sans tiers | 6 | 855 | 4 | 857 | 808 |
| VL Avec Tiers | 1072 | 2328 | 619 | 2780 | 2665 |
| VL Sans Tiers | 273 | 1012 | 177 | 1108 | 1066 |
| Piéton Avec Tiers | 449 | 705 | 277 | 877 | 853 |
| Autre Avec Tiers | 117 | 114 | 61 | 170 | 163 |
| Autre Sans Tiers | 48 | 134 | 24 | 158 | 150 |
| Total | 2785 | 7417 | 1706 | 8496 | 8174 |

## 3.4 Etape 2 : Modélisation du sous-enregistrement par capture-recapture

Lorsque deux (ou plusieurs) sources d'enregistrement d'une même population existent, il est possible avec la méthode de capture-recapture, sous certaines conditions, d'estimer le nombre de personnes échappant aux deux sources et ainsi l'effectif total de la population étudiée. La méthode de capture-recapture et son application en épidémiologie sont exposées brièvement ci-après.

### 3.4.1 Synthèse bibliographique sur capture-recapture

Trois articles (Hook and Regal 1995; IWGDMF 1995; Gallay, Nardone et al. 2002) fournissent une revue bibliographique particulièrement intéressante de la méthode de capture-recapture, et avec un point de vue épidémiologique ; nous présentons ici une brève synthèse bibliographique de son utilisation.

La méthode de capture-recapture a été mise en place en écologie animale, pour quantifier des populations de poissons, d'oiseaux… Par exemple, s'il s'agit de quantifier la population de poissons dans un étang, le principe est de faire une première capture, de marquer les individus capturés, de les relâcher, puis de faire une deuxième capture, et parmi ceux capturés, d'identifier s'il s'agit de leur première capture, ou de leur deuxième (déjà marqués). À partir des effectifs de chaque sous-groupe (capturés seulement la 1ère fois, seulement la 2ème fois, ou les deux fois), et de calculs de probabilités, sous certaines hypothèses à vérifier, on estime l'effectif de ceux jamais capturés et donc l'effectif total.

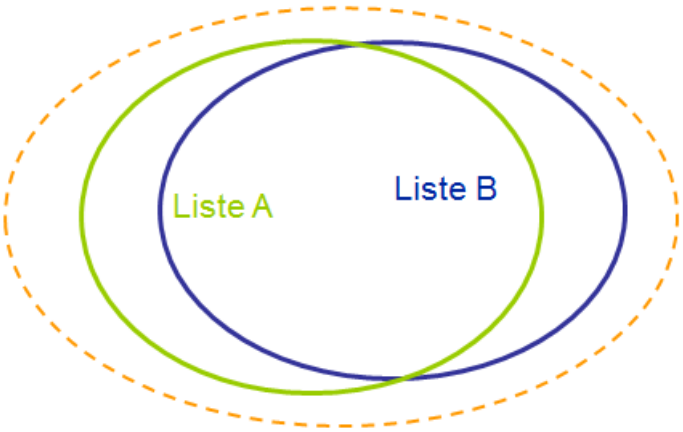


*Figure 7 : schéma de deux recensements (ou « listes »)*

Une autre façon de présenter les choses consiste à les écrire sous forme d'un tableau :

*Tableau 18: répartition des effectifs selon l'appartenance des sujets à deux sources d'enregistrement*

| | | Liste B | | |
|---|---|---|---|---|
| | | oui | non | |
| Liste A | oui | $n_{AB}$ | $n_{A\overline{B}}$ | $n_A$ |
| | non | $n_{\overline{A}B}$ | $n_{\overline{A}\overline{B}}$ | |
| | | $n_B$ | | $n$ |

Si l'on suppose que la proportion d'appartenance à la liste A, parmi la population totale, est la même que la proportion (d'appartenance à la liste A) parmi ceux appartenant à la liste B, ce qui s'écrit $n_A/n = n_{AB}/n_B$, alors on obtient l'estimateur intuitif de Petersen $\hat{n} = \frac{n_A \times n_B}{n_{AB}}$ ; il est également l'estimateur du maximum de vraisemblance.

Par transposition des captures à des listes d'enregistrement, l'approche a été utilisée dans d'autres domaines : en démographie, pour mesurer la couverture d'un recensement (Fienberg 1992; Garton, Abdalla et al. 1996), en épidémiologie, surtout depuis les années 80 (Hook and Regal 1995; IWGDMF 1995; Gallay, Nardone et al. 2002). Elle sert à estimer des effectifs des populations, en particulier des populations « cachées » (Neugebauer and Wittes 1994) : séropositifs (Frischer, Bloor et al. 1991; Abeni, Brancato et al. 1994), personnes atteintes du SIDA (Bernillon, Lievre et al. 2000), toxicomanes (Hay, McKeganey et al. 1999; Maxwell 2000; Gemmell, Millar et al. 2004), prostituées (Bloor, Leyland et al. 1991), personnes sans domicile fixe (Fisher, Turner et al. 1994)…. Elle sert aussi à estimer le taux de couverture de registres ou encore à estimer des incidences corrigées de la sous-déclaration ou du sous-enregistrement. Les deux derniers objectifs sont essentiellement observés dans les domaines suivants : diabète (IWGDMF 1995; IWGDMF 1995), malformations congénitales, cancer (Sherman 1981; Robles, Marrett et al. 1988; Hilsenbeck, Kurucz et al. 1992; Brenner, Stegmaier et al. 1994; Schouten, Straatman et al. 1994; Brenner, Stegmaier et al. 1995; Crocetti, Miccinesi et al. 2001), VIH et SIDA (Frischer, Leyland et al. 1993; Abeni, Brancato et al. 1994; Johri, Kaplan et al. 1999;

Bernillon, Lievre et al. 2000), traumatologie (Roberts and Scragg 1994; LaPorte, Dearwater et al. 1995; Johnson, Gabella et al. 1997; Rosenman, Kalush et al. 2006).

La méthode de capture-recapture est aussi utilisée dans l'épidémiologie des traumatismes de la circulation routière. Elle a été appliquée sur les sous-populations suivantes : enfants (Roberts and Scragg 1994; Jarvis, Lowe et al. 2000), adolescents et jeunes adultes (Morrison and Stone 2000), piétons et cyclistes (Dhillon, Lightstone et al. 2001), conducteurs de poids lourds (Meuleners, Lee et al. 2006), mais aussi à des zones géographiques : villes (Razzak and Luby 1998; Tercero and Andersson 2004) ou une île (la Réunion) (Aptel, Salmi et al. 1999).

L'application de la méthode de capture-recapture en épidémiologie est sujette à des prises de position très tranchées. Certains chercheurs pensent qu'il est bien mieux d'appliquer cette méthode que de ne rien faire, et que tout registre évaluant son taux de couverture doit l'utiliser. D'autres chercheurs sont franchement réfractaires et rejettent la méthode en bloc. Le sujet est polémique ; il faut dire que la méthode est victime de sa simplicité d'application : elle est parfois utilisée un peu trop vite, sans vraiment tenir compte des conditions d'application. D'autres chercheurs tentent de calmer la polémique, en discutant point par point les conditions d'application. Une liste de 17 recommandations a notamment été écrite (Hook and Regal 1999; Hook and Regal 2000) concernant l'application de la méthode, la vérification des conditions d'application, la présentation des étapes de l'analyse et des résultats. Dans le domaine de l'épidémiologie des traumatismes, une poignée de chercheurs se montre aussi très sceptique (Jarvis, Lowe et al. 2000; Morrison and Stone 2000) ; cependant, leur discussion des conditions d'applications n'apparait pas complète.

Concernant les méthodes d'estimation, il y a eu historiquement les estimateurs « simples » (nommés estimateur de Petersen, de Sekar et Deming…) limités aux situations à deux sources d'enregistrement (avec notamment l'appellation de « méthode à deux listes »), où il suffit d'une calculette pour estimer le nombre de sujets non-observés. S'est ensuite développée la modélisation explicite. Celle-ci permet de prendre facilement en compte des variables liées à la probabilité d'enregistrement, mais aussi, des éventuelles interactions entre les listes, lorsqu'il y en a trois ou plus. La modélisation la plus répandue est celle basée sur une loi de Poisson, avec un modèle log-linéaire (Cormack 1989; IWGDMF 1995; Chao, Tsay et al. 2001) ; elle est liée à l'analyse des tableaux de contingence. Une autre modélisation se base sur la loi multinomiale et le modèle logistique (Alho 1990; Tilling and Sterne 1999). Il existe aussi des développements non-paramétriques, bayésiens (Madigan and York 1997; Hook and Regal 2000) et autres qui n'ont pas été explorés ici.

Nous l'appliquons ici grâce à une modélisation explicite, par un modèle multinomial logistique. Cela nous permet de prendre en compte l'ensemble des facteurs majeurs de biais de sélection, et aussi d'estimer directement les coefficients de correction du sous-enregistrement des données des forces de l'ordre.

### 3.4.2 Conditions d'application

L'approche capture-recapture se base sur quatre conditions clés d'application (Hook and Regal 1995; IWGDMF 1995; IWGDMF 1995; Gallay, Nardone et al. 2002) :
1) population fermée,
2) indépendance d'enregistrement entre les deux sources,
3) homogénéité de la probabilité de capture par une source donnée,
4) identification parfaite des sujets communs aux deux sources.

Deux autres conditions existent (Gallay, Nardone et al. 2002) mais sont rarement explicitées, car considérées comme évidentes. Il n'est cependant pas inutile de les donner :

5) même période de temps et même zone géographique,
6) pas d'erreur de classement des cas.

**La condition 1** de population fermée, signifie qu'il n'y a pas d'entrée ou sortie entre les deux sources d'enregistrement. Sur l'exemple des poissons, cela signifie qu'il n'y pas d'apport ou de perte de poissons entre les deux pêches, ni naissances ni décès.
Sur les blessés de la route, cela signifie qu'il ne doit pas y avoir d'entrée ou de sortie entre les enregistrements des forces de l'ordre et du Registre, c'est-à-dire entre l'établissement d'un procès-verbal par les forces de l'ordre d'un côté, et la consultation dans un service hospitalier. Cette condition est vérifiée en grande partie. Peu de temps s'écoule entre la venue des forces de l'ordre sur le lieu de l'accident et la consultation à l'hôpital (au plus quelques jours pour les plus légèrement blessés). La seule possibilité de sortie concerne les personnes blessées dans le Rhône se rendant dans un hôpital non couvert par le Registre, c'est-à-dire hors du Rhône et de ses hôpitaux limitrophes. Ce seraient typiquement des blessés légers (MAIS 1) n'habitant pas le Rhône.
Combien sont-ils ? Une indication de la proportion de blessés impliqués dans un accident dans le département du Rhône mais habitant hors du Rhône est donnée dans les données des forces de l'ordre : 24 % de voitures impliquées dans un accident corporel sont immatriculées dans un département autre que le Rhône.
Parallèlement, parmi les blessés occupants de voiture recensés dans le Registre, 16 % habitent un autre département que le Rhône. S'il n'y avait aucun biais de sélection, ni du Registre ni des données des forces de l'ordre, les deux proportions devraient être égales, et éventuellement égales à la proportion dans les BAAC soit 24%. Nous pourrions supposer alors qu'il nous manque 8 % (24 %-16 %) des automobilistes blessés. Il faut faire de même pour chaque type d'usager car la mobilité interdépartementale est bien sûr liée au mode de transport. La proportion de personnes habitant hors du Rhône est très faible parmi les blessés piétons et cyclistes. Elle est forte parmi les blessés usagers de véhicules utilitaires et poids lourds. **Cela conduit à ce que les estimations finales d'effectifs de blessés, surtout légers, soient des valeurs planchers.**

**La condition 2** d'indépendance d'enregistrement entre les deux sources signifie que les enregistrements sont faits de façon indépendante. Sur l'exemple des poissons, c'est le cas par défaut. Pour les blessés de la route, ce n'est pas le cas. Il y a une dépendance positive pour les blessés graves : en effet, en présence d'un blessé supposé grave, les forces de l'ordre appellent en général les pompiers ou le SAMU. La réciproque existe, mais dans une moindre mesure. Cette corrélation porte sur l'alerte et la présence sur le lieu de l'accident des équipes des forces de l'ordre et des équipes médicales. Cette corrélation « sur le terrain » entraîne une corrélation positive sur les probabilités d'enregistrement dans les données des forces de l'ordre et dans le Registre. Il a été démontré (IWGDMF 1995) qu'une corrélation positive se traduit par une sous-estimation du nombre total de blessés. **Les effectifs et incidences estimés de blessés graves sont donc vraisemblablement des valeurs planchers.**

**La condition 3** d'homogénéité de capture signifie que tous les individus d'intérêt ont la même probabilité d'être enregistrés par une source donnée. Sur l'exemple des poissons : que tous les poissons, quelle que soit leur taille, leur âge, ont la même probabilité d'être pêchés. En épidémiologie, il est en fait fréquent que l'homogénéité de capture ne soit vérifiée qu'à l'intérieur de sous-groupes, définis par exemple par les niveaux de gravité d'une maladie. Cela peut et doit être pris en compte dans la mise en application de la méthode, soit par stratification sur la variable définissant les sous-groupes soit par inclusion de la variable dans une modélisation explicite, ce qui est le cas, ici. Pour les blessés de la route, ces variables correspondent à ce que nous appelons biais de sélection. Du côté de l'enregistrement dans les BAAC, les facteurs de biais de sélection les plus importants sont la gravité lésionnelle, le type d'usager, la présence / absence de tiers, le type de réseau, et le type de force de l'ordre (Amoros, Martin et al. 2006) ; la gravité de l'accident (accident mortel ou corporel) s'est aussi avérée jouer sur la probabilité d'enregistrement des blessés impliqués

dans ces accidents. En ce qui concerne les biais de sélection du Registre (Amoros 2007), il s'agit de la gravité lésionnelle, de la distance entre le lieu d'accident et l'hôpital le plus proche, et éventuellement du type d'usager (sous-notification des cyclistes).

**La condition 4** d'identification parfaite des sujets communs aux deux sources signifie que le chaînage entre les 2 sources d'enregistrement est parfait : on ne « loupe » aucun rapprochement entre 2 enregistrements de la même victime (i.e. il n'y a aucun « faux négatif » par rapport au chaînage), et que pour chaque paire reliée entre les 2 sources, il s'agit bien de la même victime (i.e. il n'y a aucun « faux positif »).
En pratique, cette condition est difficile à satisfaire complètement, et à évaluer. Nous avons cependant amélioré le chaînage, dans l'étape 2, en estimant le nombre de faux positifs et de faux positifs, et en corrigeant le résultat du chaînage de ces effectifs.

**La condition 5** « même période de temps et même zone géographique » parait une évidence. Dans le cas des poissons, cela signifie que les deux pêches doivent être très rapprochées dans le temps, et sur la même zone de l'étang.
La période de temps est ici 2006-2012. Le chaînage des victimes entre les données des forces de l'ordre et le Registre se fait année par année. La période de temps est la même entre les deux sources d'enregistrement.
La zone géographique est le département du Rhône, comme lieu d'accident. Du côté des données des forces de l'ordre, le département où s'est produit l'accident est toujours renseigné, et c'est cette variable qui sert à identifier et à sélectionner les accidents ayant eu lieu dans le département du Rhône. Du côté du Registre, le critère d'inclusion géographique est aussi le département du Rhône, en tant que lieu d'accident.
Le lieu de l'accident est assez souvent non renseigné parmi les victimes du Registre. On peut penser que par défaut les services hospitaliers dans le département du Rhône incluent soit toutes les personnes qui se présentent suite à un accident de la route, soit font une sélection sur « accident dans le Rhône » mais sans demander plus de précision ; en effet, la pertinence de la variable « lieu de l'accident » n'est pas flagrante pour les soignants et il est donc compréhensible que cette variable soit peu renseignée. Le personnel du Registre envoie un courrier aux victimes du Registre lorsque cette information est manquante mais elle le reste chez 25,6 % des blessés (nom de la commune non renseigné).
Lors de l'enquête auprès des blessés identifiés seulement dans le Registre (Amoros 2007), il est apparu que certains accidents, parmi ceux où le lieu n'était pas renseigné, avaient en fait eu lieu hors du Rhône. Cela correspond à 2 % des répondants, parmi les blessés par accident en 2001, identifiés seulement dans le Registre.
La condition n°5 de l'approche capture-recapture implique d'exclure ces victimes blessées dans un accident hors Rhône. Elles sont cependant difficilement identifiables.
En pratique nous avons choisis de procéder de la manière suivante : nous généralisons le pourcentage de tels blessés (parmi les répondants) à l'ensemble des blessés identifiés dans le Registre seulement, et à l'ensemble des années de la période étudiée. Nous appliquons ainsi ces 2 % aux blessés identifiés seulement dans le Registre, année par année ; cela nous donne un effectif de blessés à exclure. Nous tirons au sort ces blessés, parmi les blessés identifiés seulement dans le Registre ET dont le lieu de l'accident est manquant.

**La condition 6** « pas d'erreur de classement des cas » signifie que les cas qui nous intéressent (poissons de telle espèce, ou blessés de la route) sont parfaitement identifiables et identifiés, ou autrement dit, qu'ils sont des vrais cas (qu'il n'y ait ni faux négatifs ni faux positifs, par rapport à la définition d'un cas). Ici, il s'agit de s'assurer que les personnes vérifient les deux critères : « blessées » et « lors d'un accident de la circulation routière ».
Du côté des BAAC, l'étude des blessés identifiés dans les BAAC seulement (Amoros 2007) a mis en évidence que certaines personnes classées « blessées » ne l'étaient pas : typiquement il s'agit de

personnes envoyées dans un service hospitalier pour observation (femmes enceintes, nourrissons par exemple), ou de personnes en état de choc psychologique, mais non blessées physiquement (la classification AIS ne retient pas le choc psychologique comme une blessure), ou encore de personnes déclarées « blessées » par précaution vis-à-vis des assurances ou d'éventuelle procédure judiciaire mais qui se révèlent indemnes cliniquement.
Afin de vérifier la condition n°6 de l'approche capture-recapture, il nous faut exclure ces personnes des données des forces de l'ordre. Nous procédons de la même façon que pour satisfaire à la condition n°5 : nous généralisons la proportion observée (4,1 %) sur les blessés de cette étude (ceux avec PV exploitable), à l'ensemble des blessés identifiés dans les BAAC seulement, et à l'ensemble de la période 2006-2012. Nous tirons au sort les personnes à exclure (année par année), parmi les blessés identifiés seulement dans les BAAC et parmi ceux classés blessés légers (au sens BAAC). Nous supposons en effet que ceux classés blessés graves sont réellement blessés et que l'erreur ne porte que sur la frontière blessé léger / indemne.
En ce qui concerne le critère « accident de la circulation routière », il parait peu probable que d'autres types d'accidents (professionnel ou de loisirs ou de la vie courante, exclusivement) soient classés comme accidents de la circulation routière par les forces de l'ordre. Il arrive que des accidents de la route soient des suicides. Il est difficile de les identifier.

Du côté du Registre, les personnes incluses sont assurément blessées puisqu'un des critères d'inclusion du Registre est « personne présentant au moins une lésion au sens de la classification AIS (gravité AIS 1 au minimum ou AIS 9 indéterminé, correspondant à une ou plusieurs lésions mal définies) ».
En ce qui concerne le critère « accident de la circulation routière » l'enquête sur les blessés identifiés dans le Registre seulement (Amoros 2007) a montré que quelques personnes s'avéraient ne pas correspondre à cette définition (huit parmi 547, soit 1,2 % en tenant compte du plan de sondage). Il s'agit d'accidents professionnels ou de la vie courante ou de loisirs, exclusivement, ou encore d'accidents avec un moyen de transport non routier.
Afin de vérifier la condition n°6 de l'approche capture-recapture, nous généralisons le pourcentage de « hors accident de la route » à l'ensemble des blessés identifiés dans le Registre seulement, et à la période 2006-2012. Nous tirons au sort les personnes à exclure (année par année) parmi les blessés identifiés dans le Registre seulement.

Les critères de définition des données des forces de l'ordre, qui sont ceux utilisés par le ministère chargé des transports pour communiquer les bilans officiels en termes de sécurité routière, sont plus restrictifs que ceux du Registre, et ont été appliqués aux victimes du Registre. Nous rappelons que, pour cette raison, les usagers de rollers, skateboard, trottinette se blessant seuls ou contre un piéton, inclus dans le Registre, sont exclus de la projection. Pour la même raison, les victimes du Registre d'accidents clairement identifiés comme hors réseau routier sont exclues de la projection.

### 3.4.3 Construction du modèle (logit multinomial)

Pour modéliser l'approche de capture-recapture, nous utilisons un modèle logistique multinomial ; il permet d'inclure facilement les variables liées aux probabilités d'enregistrement. Nous nous sommes fortement appuyés sur un article de Tilling et Sterne (Tilling and Sterne 1999). La présentation ci-dessous reprend une partie de leur article.

#### *3.4.3.1 Sous l'hypothèse d'indépendance des deux sources*

On formalise l'enregistrement par les deux sources BAAC et Registre avec le tableau suivant :

| | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | | | $N_B$ |
| | non | | | |
| | | $N_R$ | | N ? |

Avec

$N_B$=Nombre de blessés dans les BAAC (B=BAAC)
$N_R$=Nombre de blessés dans le Registre (R=Registre)

Le modèle s'écrit de la façon suivante :

$\Pi_B$ = probabilité d'être enregistré dans les BAAC
$\Pi_R$ = probabilité d'être enregistré dans le Registre

Dans le tableau croisé, cela correspond aux probabilités sur les marges :

| | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | | | $\Pi_B$ |
| | non | | $\Pi_0$ | |
| | | $\Pi_R$ | | |

On note $l_B$ et $l_R$ les logit de ces probabilités

$l_B = \text{logit}(\Pi_B) = ln\left(\frac{\pi_B}{1-\pi_B}\right)$

$l_R = \text{logit}(\Pi_R) = ln\left(\frac{\pi_R}{1-\pi_R}\right)$

Sous l'hypothèse d'indépendance des deux sources, la probabilité pour tout individu d'être manqué par les deux recensements vaut :

$$\Pi_0=(1-\Pi_B)\,(1-\Pi_R)=\frac{1}{(1+\exp(l_B))\times(1+\exp(l_R))}$$

Une variable indicatrice $q_i$ est définie ; elle identifie si l'individu i est capturé par aucune source ($q_i$=0), par la source BAAC seulement ($q_i$=1), par la source Registre seulement ($q_i$=2) ou par les deux sources ($q_i$=3). En termes d'effectifs, cela s'écrit, dans le tableau croisé, de la façon suivante :

| | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | $N_3$ | $N_1$ | $N_B$ |
| | non | $N_2$ | $\widehat{N0}$ | |
| | | $N_R$ | | $\widehat{N}$ |

$N = N_1 + N_2 + N_3$ (en termes d'observé)
$\widehat{N} = N + \widehat{N_0}$ (le « $\widehat{\ }$ » dénote l'estimateur, et l'estimation)
avec $\widehat{N}$ estimateur du nombre réel des blessés, $\widehat{N_0}$, estimateur du nombre réel de blessés manqués

On conditionne sur les individus observés (q ≠0) afin d'estimer les paramètres du modèle, pour ensuite pouvoir estimer la taille de la population étudiée. On note k=1,2,3 les sous-ensembles exclusifs « BAAC seulement », « Registre seulement » et « intersection BAAC-Registre ». On note $p_{ik}$ avec k=1, 2, 3 la probabilité que $q_i$= k, sachant que l'individu i est observé.

Par hypothèse d'homogénéité de capture, les $p_{ik}$ sont identiques pour tous les individus i et donc $p_{ik}$ est remplacé par $p_k$. On a les probabilités conditionnelles suivantes :

$p_1$= P(être dans BAAC seulement) / P(être observé)
$p_2$ = P(être dans Registre seulement) / P(être observé)
$p_3$ = P(être dans l'intersection BAAC - Registre) / P(être observé)

Cela correspond à l'écriture ci-dessous dans le tableau croisé :

| | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | $p_3$ | $p_1$ | $\Pi_B$ |
| | non | $p_2$ | $\Pi_0$ | |
| | | $\Pi_R$ | | |

On a donc, par hypothèse d'indépendance des deux sources :

$$p_1 = \frac{\Pi_B(1-\Pi_R)}{\Pi_B(1-\Pi_R)+(1-\Pi_B)\Pi_R+\Pi_B\Pi_R} = \frac{\exp(l_B)}{\exp(l_B)+\exp(l_R)+\exp(l_B+l_R)}$$

De la même façon :

$$p_2 = \frac{\exp(l_R)}{\exp(l_B)+\exp(l_R)+\exp(l_B+l_R)}, \text{ et}$$

$$p_3 = \frac{\exp(l_B+l_R)}{\exp(l_B)+\exp(l_R)+\exp(l_B+l_R)}$$

On formule un modèle logit sur ces probabilités conditionnelles $p_k$. La variable à expliquer est une variable catégorielle nominale ; on utilise un modèle logit généralisé, en prenant k=1 i.e. « BAAC seulement » comme référence. Le modèle s'écrit :

$\text{logit}(p_k)=\log(p_k/p_1) = ß_k$ , k=2,3

et en posant $ß_1$=0 (correspondant à « BAAC seulement » pris comme référence).

En estimant le modèle à partir des données, on obtiendra une estimation des paramètres $ß_k$. Or, les paramètres $ß_k$ sont reliés à $l_1$ et $l_2$ de la façon suivante :

$$\beta_2 = \log\left(\frac{p_2}{p_1}\right) = \log\left(\frac{\exp(l_R)/\left(\exp(l_B)+\exp(l_R)+\exp(l_B+l_R)\right)}{\exp(l_B)/\left(\exp(l_B)+\exp(l_R)+\exp(l_R+l_B)\right)}\right) = \log\left(\frac{\exp(l_R)}{\exp(l_B)}\right) = l_R - l_B$$

$$\beta_3 = \log\left(\frac{p_3}{p_1}\right) = \log\left(\frac{\exp(l_B+l_R)}{\exp(l_B)}\right) = l_B + l_R - l_B = l_R$$

et donc réciproquement :

$$l_R = \beta_3$$
$$l_B = \beta_3 - \beta_2$$

On estime le modèle sur les observations, et on obtient les estimations $\hat{\beta}_2$ et $\hat{\beta}_3$

En remplaçant les paramètres $l_B$ et $l_R$ par leurs estimations, fonctions des estimations des paramètres $ß_k$, on obtient une estimation de $\Pi_0$ :

$$\hat{\Pi}_0 = \frac{1}{(1+\exp(\hat{l}_B))\times(1+\exp(\hat{l}_R))}$$
$$= \frac{1}{(1+\exp(\hat{\beta}_3-\hat{\beta}_2))\times(1+\exp(\hat{\beta}_3))}$$
$$= \frac{\exp(\hat{\beta}_2)}{(\exp(\hat{\beta}_2)+\exp(\hat{\beta}_3))\times(1+\exp(\hat{\beta}_3))}$$

On peut alors estimer le nombre total N. En effet :

$$\hat{\Pi}_0 = \frac{\hat{N}_0}{\hat{N}} \Leftrightarrow$$
$$\hat{\Pi}_0 = \frac{\hat{N}-(N_1+N_2+N_3)}{\hat{N}} \Leftrightarrow$$
$$1-\hat{\Pi}_0 = \frac{(N_1+N_2+N_3)}{\hat{N}} \Leftrightarrow$$
$$\hat{N} = \frac{(N_1+N_2+N_3)}{1-\hat{\Pi}_0}$$

La modélisation logit multinomiale nous donne aussi une estimation des probabilités $\Pi_B$ et $\Pi_R$ :

$$\hat{\Pi}_B = \frac{1}{1+\exp(-\hat{l}_B)} = \frac{1}{1+\exp(\hat{\beta}_2-\hat{\beta}_3)} \text{ et}$$
$$\hat{\Pi}_R = \frac{1}{1+\exp(-\hat{l}_R)} = \frac{1}{1+\exp(-\hat{\beta}_3)}$$

L'inverse de $\Pi_B$ correspond au coefficient de correction du sous-enregistrement des données des forces de l'ordre, facteur multiplicatif à appliquer aux effectifs des BAAC.

$\hat{C}_B = \frac{1}{\hat{\Pi}_B} = 1+\exp(\hat{\beta}_2-\hat{\beta}_3)$, coefficient de correction du sous-enregistrement des données des forces de l'ordre.

On peut aussi estimer $C_R$, coefficient de correction du sous-enregistrement du Registre :

$$\hat{C}_R = \frac{1}{\hat{\Pi}_R} = 1+\exp(-\hat{\beta}_3),$$

Enfin, on peut estimer $C_3$, coefficient de correction du sous-enregistrement de l'intersection BAAC Rhône et Registre :

$$\hat{C}_3 = \frac{1}{\hat{p}_3} = \left[1+\exp(\hat{\beta}_2)+\exp(\hat{\beta}_3)\right] \Big/ \exp(\hat{\beta}_3)$$

Cela est utile pour toute analyse restreinte à ces données, qui ont l'avantage de combiner caractéristiques accidentologiques et caractéristiques médicales (par exemple : étudier les blessures des automobilistes sur telle région du corps, en fonction du lieu d'impact sur le véhicule)

#### *3.4.3.2 Sous l'hypothèse de dépendance des deux sources*

Nous considérons maintenant la situation où les deux sources sont dépendantes, dans le sens où la probabilité d'enregistrement dans l'une et dans l'autre sont liées à une même caractéristique (notée X) des individus à enregistrer. L'hypothèse d'indépendance des deux sources est alors posée, dans des catégories définies par X, c'est-à-dire conditionnellement à X.

Toujours en notant i l'individu, les probabilités d'enregistrement dans les sources BAAC (B) et Registre (R) sont notées $\Pi_{Bi}$ et $\Pi_{Ri}$.
On modélise leur lien à la caractéristique X (quantitative) de la façon suivante :

$$\text{logit}(\Pi_{Bi}) = l_B + k_B X_i$$

$$\text{logit}(\Pi_{Ri}) = l_R + k_R X_i$$

Rien ne contraint $k_B$ et $k_R$ à être de valeurs égales.

On prend comme caractéristique une variable qualitative telle que la gravité lésionnelle en deux classes : BL, BH (gravité BAAC, BL= blessé léger, BH=blessé hospitalisé), que l'on note m (m=1, 2). Pour une catégorie m donnée, le tableau croisé BAAC x Registre s'écrit :

| catégorie m | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | | | $\Pi_{Bm}$ |
| | non | | $\Pi_{0m}$ | |
| | | $\Pi_{Rm}$ | | |

Pour tout individu appartenant à la catégorie m, on pose :

$$\text{logit}(\Pi_{Bm}) = l_B + k_{Bm}$$

$$\text{logit}(\Pi_{Rm}) = l_R + k_{Rm}$$

La probabilité d'un individu i, de la catégorie m, d'être manqué (par les deux sources) est donnée, sous l'hypothèse d'indépendance des 2 sources à l'intérieur de la catégorie m, par :

$\Pi_{0i} = (1 - \Pi_{Bi})\,(1 - \Pi_{Ri})$

Cela s'écrit, pour tout individu i dans la catégorie m :

$\Pi_{0m} = (1 - \Pi_{Bm})\,(1 - \Pi_{Rm})$

$$\Pi_{0m} = \frac{1}{(1+\exp(l_B + k_{Bm}))} \times \frac{1}{(1+\exp(l_R + k_{Rm}))}$$

On définit la variable indicatrice $q_i$ comme précédemment : elle identifie si l'individu i est capturé par aucune source ($q_i$=0), par la source BAAC seulement ($q_i$=1), par la source Registre seulement ($q_i$=2) ou par les deux sources ($q_i$=3). En termes d'effectifs, le tableau croisé s'écrit, pour une catégorie m :

| catégorie m | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | $N_{3m}$ | $N_{1m}$ | $N_{Bm}$ |
| | non | $N_{2m}$ | $\widehat{N_{0m}}$ | |
| | | $N_{Rm}$ | | $\widehat{Nm}$ |

Avec

$N_m = N_{1m} + N_{2m} + N_{3m}$ (en termes d'observé)

$$\widehat{N_m} = N_m + \widehat{N_{0m}}$$

avec $\widehat{N_m}$ estimateur du vrai effectif des blessés, et $\widehat{N_{0m}}$, estimateur du vrai nombre de blessés manqués.

On a : $\widehat{N} = \sum_{m=1}^{2} \widehat{N}_m$

De façon identique, on conditionne sur les individus observés (q ≠0) afin d'estimer les paramètres du modèle, pour ensuite pouvoir estimer la taille de la population étudiée. On note toujours k=1,2,3 les sous-ensembles exclusifs « BAAC seulement », « Registre seulement » et « intersection BAAC-Registre ». On note $p_{ik}$ avec k=1, 2, 3 la probabilité que $q_i$= k, sachant que l'individu i est observé.

Par hypothèse d'homogénéité de capture à l'intérieur de la catégorie m, les $p_{ik}$ sont identiques pour tous les individus i et donc $p_{ik}$ est remplacé par $p_{km}$. On a les probabilités conditionnelles suivantes :

$p_{1m}$= P(être dans BAAC seulement) / P(être observé)
$p_{2m}$ = P(être dans Registre seulement) / P(être observé)
$p_{3m}$ = P(être dans l'intersection BAAC - Registre) / P(être observé)

Cela s'écrit, pour la catégorie m :

| catégorie m | | Registre (R) | | |
|---|---|---|---|---|
| | | oui | non | |
| BAAC (B) | oui | $p_{3m}$ | $p_{1m}$ | $\Pi_B$ |
| | non | $p_{2m}$ | $\Pi_{0m}$ | |
| | | $\Pi_R$ | | |

Pour tout individu i appartenant à la catégorie m, on a, par hypothèse d'indépendance des 2 sources, à catégorie m fixée :

$$p_{1m} = \frac{\Pi_{Bm}(1-\Pi_{Rm})}{\Pi_{Bm}(1-\Pi_{Rm})+(1-\Pi_{Bm})\Pi_{Rm}+\Pi_{Bm}\Pi_{Rm}} = \frac{\exp(l_B + k_{Bm})}{\exp(l_B + k_{Bm})+\exp(l_R + k_{Rm})+\exp(l_B + k_{Bm} + l_R + k_{Rm})}$$

De la même façon :

$$p_{2m} = \frac{\exp(l_R + k_{Rm})}{\exp(l_B + k_{Bm})+\exp(l_R + k_{Rm})+\exp(l_B + k_{Bm} + l_R + k_{Rm})}$$

$$p_{3m} = \frac{\exp(l_B + k_{Bm} + l_R + k_{Rm})}{\exp(l_B + k_{Bm})+\exp(l_R + k_{Rm})+\exp(l_B + k_{Bm} + l_R + k_{Rm})}$$

Ces probabilités sont aussi formalisées par un modèle multinomial généralisé, qui s'étend sans problème à l'inclusion d'une variable X :

$\text{logit}(p_{ki}) = \log\left(\frac{p_{ki}}{p_{1i}}\right) = \beta_k + \theta_k X_i$, avec k=2,3, et k=1 la catégorie de référence (ß$_1$=0, et Θ$_1$=0)

Ou, si l'on inclut une variable catégorielle, telle que la gravité BAAC en 2 catégories (m=1, 2), on écrit, pour tout individu i appartenant à la catégorie m :

$\text{logit}(p_{km}) = \log\left(\frac{p_{km}}{p_{1m}}\right) = \beta_k + \theta_{km}$, avec k=2,3, et k=1 la catégorie de référence (ß$_1$=0, et Θ$_{1m}$=0)

En ajustant le modèle, on va obtenir une estimation des paramètres ß$_k$ et Θ$_{km}$.

Or, ils sont reliés aux paramètres l et k de la façon suivante :

$$\beta_2 + \theta_{2m} = \log\left(\frac{p_{2m}}{p_{1m}}\right)$$
$$= \log\left(\frac{\exp(l_R + k_{Rm})/\left(\exp(l_B + k_{Bm}) + \exp(l_R + k_{Rm}) + \exp(l_B + k_{Bm} + l_R + k_{Rm})\right)}{\exp(l_B + k_{Bm})/\left(\exp(l_B + k_{Bm}) + \exp(l_R + k_{Rm}) + \exp(l_B + k_{Bm} + l_R + k_{Rm})\right)}\right)$$
$$= \log\left(\frac{\exp(l_R + k_{Rm})}{\exp(l_B + k_{Bm})}\right)$$
$$= (l_R + k_{Rm}) - (l_B + k_{Bm})$$

$$\beta_3 + \theta_{3m} = \log\left(\frac{p_{3m}}{p_{1m}}\right)$$
$$= \log\left(\frac{\exp(l_B + k_{Bm} + l_R + k_{Rm})/\left(\exp(l_B + k_{Bm}) + \exp(l_R + k_{Rm}) + \exp(l_B + k_{Bm} + l_R + k_{Rm})\right)}{\exp(l_B + k_{Bm})/\left(\exp(l_B + k_{Bm}) + \exp(l_R + k_{Rm}) + \exp(l_B + k_{Bm} + l_R + k_{Rm})\right)}\right)$$
$$= \log\left(\frac{\exp(l_B + k_{Bm} + l_R + k_{Rm})}{\exp(l_B + k_{Bm})}\right)$$
$$= l_R + k_{Rm}$$

Pour la catégorie m de référence, on a ß$_1$=0, et Θ$_{1m}$=0, et donc on en déduit :

$$\beta_2 = l_R - l_B$$
$$\beta_3 = l_R$$

et ensuite :

$$\theta_{2m} = k_{Rm} - k_{Bm}$$
$$\theta_{3m} = k_{Rm}$$

et donc, réciproquement :

$$l_B = \beta_3 - \beta_2$$
$$l_R = \beta_3$$
$$k_{Bm} = \theta_{3m} - \theta_{2m}$$
$$k_{Rm} = \theta_{3m}$$

On ajuste le modèle, on remplace alors les paramètres ß et Θ par leurs estimations, et on en déduit une estimation des Π$_{0m}$ :

$$\hat{\Pi}_{0m} = \frac{1}{(1+\exp(\hat{l}_B + \hat{k}_{Bm}))\times(1+\exp(\hat{l}_R + \hat{k}_{Rm}))}$$
$$= \frac{1}{(1+\exp(\hat{\beta}_3 - \hat{\beta}_2 + \hat{\theta}_{3m} - \hat{\theta}_{2m}))\times(1+\exp(\hat{\beta}_3 + \hat{\theta}_{3m}))}$$

On peut alors estimer le nombre total par catégorie m, $\widehat{N_m}$.

En effet :

$$\hat{\Pi}_{0m} = \frac{\hat{N}_{0m}}{\hat{N}_m} \Leftrightarrow$$

$$\hat{\Pi}_{0m} = \frac{\hat{N}_m - (N_{1m} + N_{2m} + N_{3m})}{\hat{N}_m} \Leftrightarrow$$

$$1 - \hat{\Pi}_{0m} = \frac{(N_{1m} + N_{2m} + N_{3m})}{\hat{N}_m} \Leftrightarrow$$

$$\hat{N}_m = \frac{(N_{1m} + N_{2m} + N_{3m})}{1 - \hat{\Pi}_{0m}}$$

On peut en déduire le nombre total, en sommant sur les catégories m :

$$\hat{N} = \sum_{m=1}^{2} \hat{N}_m$$

On estime aussi les probabilités non conditionnelles aux observations, d'être enregistré dans les BAAC ou dans le Registre :

$$\hat{\Pi}_{Bm} = \frac{1}{1 + \exp(-\hat{l}_B - \hat{k}_{Bm})} = \frac{1}{1 + \exp(\hat{\beta}_2 - \hat{\beta}_3 + \hat{\theta}_{2m} - \hat{\theta}_{3m})} \text{ et}$$

$$\hat{\Pi}_{Rm} = \frac{1}{1 + \exp(-\hat{l}_R - \hat{k}_{Rm})} = \frac{1}{1 + \exp(-\hat{\beta}_3 - \hat{\theta}_{3m})}$$

En prenant l'inverse de la probabilité non conditionnelle d'être enregistré dans les BAAC, on obtient le coefficient de correction du sous-enregistrement des BAAC, pour la catégorie m :

$$\hat{C}_{Bm} = \frac{1}{\hat{\Pi}_{Bm}} = 1 + \exp(\hat{\beta}_2 - \hat{\beta}_3 + \hat{\theta}_{2m} - \hat{\theta}_{3m}) = 1 + \exp(\hat{\beta}_2 + \hat{\theta}_{2m} - \hat{\beta}_3 - \hat{\theta}_{3m}),$$

On peut aussi estimer $C_{Rm}$, coefficient de correction du sous-enregistrement du Registre.

$$\hat{C}_{Rm} = \frac{1}{\hat{\Pi}_{Rm}} = 1 + \exp(-\hat{\beta}_3 - \hat{\theta}_{3m}),$$

Enfin, on peut estimer les coefficients de correction du sous-enregistrement des données de l'intersection BAAC du Rhône et Registre :

$$\hat{C}_3 = \frac{1}{\hat{p}_3} = \left[1 + \exp(\hat{\beta}_2 + \hat{\theta}_{2m}) + \exp(\hat{\beta}_3 + \hat{\theta}_{3m})\right] \Big/ \exp(\hat{\beta}_3 + \hat{\theta}_{3m})$$

On généralise sans problème à plusieurs caractéristiques liées aux probabilités d'enregistrement. C'est-à-dire, si la probabilité d'enregistrement est fonction de plusieurs variables (gravité, type d'usager, absence ou présence de tiers,..), on peut prendre cela en compte. Le modèle multinomial généralisé s'écrit alors, avec l'inclusion de variables $X_s$ (qui peuvent être continues ou catégorielles)

$$\text{logit}(p_{ki}) = \log\left(\frac{p_{ki}}{p_{1i}}\right) = \beta_k + \sum_s \theta_{ks} X_{is}$$

pour k=2,3, avec k=1 la source de référence.

NB : les coefficients de correction du sous-enregistrement sont ici indicés par m ; m identifie les catégories définies par la combinaison des variables incluses dans le modèle. Dans la suite du

document, ils sont indifféremment indicés par m ou j, la signification ne changeant pas : cela indice les variables qui jouent sur la probabilité d'enregistrement des blessés dans les BAAC

Nous utilisons le logiciel SAS, version 9.3, procédure LOGISTIC et l'option glogit.

#### 3.4.4 Modèle d'enregistrement des blessés sur 2006-2012

Il s'agit cii de choisir les variables à inclure dans le modèle.

En premier lieu, les probabilités d'enregistrement dans les BAAC et dans le Registre ne sont pas indépendantes ; elles sont positivement corrélées pour les blessés graves : sur le terrain, pour les accidents graves, les forces de l'ordre alertent les secours médicaux s'ils ne sont pas déjà présents ; la réciproque existe sans doute dans une moindre mesure. En d'autres termes, les BAAC et le Registre sont dépendants à travers la caractéristique de gravité lésionnelle. Cette variable doit donc être incluse dans le modèle.

Par ailleurs, la modélisation permet de prendre en compte la restriction d'homogénéité de capture à des sous-groupes : cela consiste à inclure dans le modèle les variables définissant ces sous-groupes. Il s'agit donc ici des facteurs de sous-enregistrement. Du côté des BAAC, les facteurs liés à un fort biais de sélection sont la gravité lésionnelle, la gravité de l'accident (mortel /corporel), le type d'usager, la présence/absence de tiers, le type de réseau, le type de force de l'ordre, et l'année pour un type de force de l'ordre (Amoros 2007). Du côté du Registre, il s'agit de la gravité lésionnelle, éventuellement du type d'usager, et de la distance à l'hôpital le plus proche (disposant d'un service d'urgences) (Amoros 2007). Cette variable n'est pas introduite car elle n'est pas directement disponible dans les BAAC (il est théoriquement possible de la déterminer à partir des lieux d'accident et des adresses d'hôpitaux disposant d'un service d'urgence mais il n'est pas envisageable en pratique de la déterminer pour la totalité des blessés des BAAC nationaux).

Nous avons considéré l'inclusion de la variable année en quantitatif : cela permet de prendre en compte une éventuelle évolution (linéaire) du phénomène de sous-enregistrement. Nous avons choisi de ne pas inclure la variable année en qualitatif dans le modèle car nous craignons un sur-ajustement aux éventuelles irrégularités des pratiques d'enregistrement dans le département du Rhône. Il s'avère que la variable année en quantitatif est seulement significative pour la zone gendarmerie.

**Valeurs manquantes (des variables incluses dans le modèle) :**
Étant donné que l'approche capture-recapture vise à estimer le nombre exhaustif des sujets d'intérêt, il n'est pas question d'exclure des observations ; les observations avec valeur manquante sont donc traitées par de l'imputation simple. Les valeurs manquantes pour âge, sexe et type d'usager ont été imputées en fonction des deux autres variables ou d'une seule autre (imputation simple dans les deux cas).
Les variables type de réseau et force de l'ordre présentent une proportion non négligeable de valeurs manquantes parmi les observations du Registre : 21,2 % et 22,5 % respectivement. Le type de réseau a été imputé en fonction du type d'usager et de la force de l'ordre en charge du secteur. Réciproquement, la force de l'ordre a été imputée en fonction du réseau et du type d'usager.

Le tableau ci-après donne les paramètres du modèle retenu sur 2006-2012. Les odds ratios sont ceux de la probabilité d'être enregistré dans « Registre seulement » (ou dans l'intersection BAAC-Registre) pour une caractéristique donnée versus une caractéristique de référence.

*Tableau 19 : modèle capture-recapture sur 2006-2012 (n=57219 blessés),; OR=odds-ratio par rapport à la catégorie de référence : source=BAAC seulement*

| variable | source | beta | OR | IC à 95% | |
|---|---|---|---|---|---|
| ordonnée à l'origine | registre seulement | 1,08 | 2,95 | 2,49 | 3,50 |
| ordonnée à l'origine | intersection | 1,19 | 3,29 | 2,77 | 3,91 |
| indicatrice gendarmerie - année (p<0,0001) | | | | | |
| année en zone gendarmerie | registre seulement | 0,05 | 1,05 | 1,02 | 1,09 |
| année en zone gendarmerie | intersection | 0,00 | 1,00 | 0,96 | 1,04 |
| type de réseau (p<0,0001) | | | | | |
| autoroutes | registre seulement | -0,88 | 0,41 | 0,35 | 0,49 |
| autoroutes | intersection | 0,20 | 1,22 | 1,02 | 1,45 |
| routes nationales et départementales | registre seulement | -0,78 | 0,46 | 0,41 | 0,51 |
| routes nationales et départementales | intersection | 0,16 | 1,17 | 1,05 | 1,31 |
| voies communales | registre seulement | | 1,00 | | |
| voies communales | intersection | | 1,00 | | |
| autres | registre seulement | -0,49 | 0,62 | 0,49 | 0,76 |
| autres | intersection | -0,20 | 0,82 | 0,66 | 1,02 |
| force de l'ordre x gravité x type d'accident (p<0,0001) | | | | | |
| Zone CRS, blessés hospitalisés, accident corporel | registre seulement | -0,81 | 0,44 | 0,33 | 0,59 |
| Zone CRS, blessés hospitalisés, accident corporel | intersection | -0,21 | 0,81 | 0,63 | 1,03 |
| Zone CRS, blessés hospitalisés, accident mortel | registre seulement | 1,07 | 2,90 | 0,33 | 25,58 |
| Zone CRS, blessés hospitalisés, accident mortel | intersection | 1,47 | 4,34 | 0,57 | 32,86 |
| Zone CRS, blessés non hospitalisés, accident corporel | registre seulement | 0,41 | 1,51 | 1,24 | 1,82 |
| Zone CRS, blessés non hospitalisés, accident corporel | intersection | -0,64 | 0,53 | 0,44 | 0,64 |
| Zone CRS, blessés non hospitalisés, accident mortel | registre seulement | -0,51 | 0,60 | 0,19 | 1,90 |
| Zone CRS, blessés non hospitalisés, accident mortel | intersection | -0,32 | 0,73 | 0,32 | 1,67 |
| zone gendarmerie, blessés hospitalisés, acc corporel | registre seulement | | 1,00 | | |
| zone gendarmerie, blessés hospitalisés, acc corporel | intersection | | 1,00 | | |
| zone gendarmerie, blessés hospitalisés, acc mortel | registre seulement | -1,46 | 0,23 | 0,09 | 0,62 |
| zone gendarmerie, blessés hospitalisés, acc mortel | intersection | 0,28 | 1,32 | 0,69 | 2,54 |
| zone gendarmerie, blessés non hospitalisés, acc corporel | registre seulement | 2,22 | 9,19 | 7,95 | 10,62 |
| zone gendarmerie, blessés non hospitalisés, acc corporel | intersection | -0,71 | 0,49 | 0,42 | 0,58 |
| zone gendarmerie, blessés non hospitalisés, acc mortel | registre seulement | -0,31 | 0,73 | 0,35 | 1,52 |
| zone gendarmerie, blessés non hospitalisés, acc mortel | intersection | -0,22 | 0,80 | 0,43 | 1,48 |
| Zone police, blessés hospitalisés, accident corporel | registre seulement | -0,52 | 0,59 | 0,49 | 0,72 |
| Zone police, blessés hospitalisés, accident corporel | intersection | 0,13 | 1,14 | 0,95 | 1,37 |
| zone police, blessés hospitalisés, accident mortel | registre seulement | -1,23 | 0,29 | 0,08 | 1,04 |
| zone police, blessés hospitalisés, accident mortel | intersection | 0,38 | 1,46 | 0,55 | 3,86 |
| zone police, blessés non hospitalisés, accident corporel | registre seulement | 0,63 | 1,87 | 1,59 | 2,21 |
| zone police, blessés non hospitalisés, accident corporel | intersection | -0,45 | 0,64 | 0,54 | 0,75 |
| zone police, blessés non hospitalisés, accident mortel | registre seulement | -1,52 | 0,22 | 0,08 | 0,63 |
| zone police, blessés non hospitalisés, accident mortel | intersection | -0,55 | 0,58 | 0,26 | 1,31 |
| type d'usager x tiers (p<0,0001) | | | | | |
| usagers de 2RM, avec tiers | registre seulement | -0,40 | 0,67 | 0,61 | 0,73 |
| usagers de 2RM, avec tiers | intersection | 0,29 | 1,34 | 1,22 | 1,47 |
| usagers de 2RM, sans tiers | registre seulement | 1,98 | 7,22 | 6,15 | 8,48 |
| usagers de 2RM, sans tiers | intersection | 0,14 | 1,15 | 0,96 | 1,37 |
| cyclistes, avec tiers | registre seulement | 0,03 | 1,03 | 0,89 | 1,19 |
| cyclistes, avec tiers | intersection | 0,17 | 1,19 | 1,02 | 1,39 |
| cyclistes, sans tiers | registre seulement | 4,42 | 83,21 | 47,08 | 147,07 |
| cyclistes, sans tiers | intersection | -0,05 | 0,95 | 0,48 | 1,86 |
| automobilistes, avec tiers | registre seulement | | 1,00 | | |
| automobilistes, avec tiers | intersection | | 1,00 | | |
| automobilistes, sans tiers | registre seulement | 0,95 | 2,60 | 2,32 | 2,90 |
| automobilistes, sans tiers | intersection | 0,21 | 1,23 | 1,10 | 1,39 |
| piétons, avec tiers | registre seulement | -0,38 | 0,68 | 0,62 | 0,75 |
| piétons, avec tiers | intersection | 0,07 | 1,07 | 0,97 | 1,19 |
| autres, avec tiers | registre seulement | -1,37 | 0,25 | 0,21 | 0,30 |
| autres, avec tiers | intersection | -0,25 | 0,78 | 0,67 | 0,91 |
| autres, sans tiers | registre seulement | 0,38 | 1,46 | 1,18 | 1,80 |
| autres, sans tiers | intersection | -0,42 | 0,66 | 0,52 | 0,83 |

L'interprétation d'un odds ratio est complexe. L'odds ratio est un rapport de cotes, c'est à-dire le ratio entre 2 cotes, au sens des cotes dans les courses de chevaux : la cote est la probabilité de gagner divisée par la probabilité de perdre. Ici la cote est la probabilité d'être enregistré dans le Registre seulement contre la probabilité d'être enregistré dans BAAC seulement.

Et le ratio entre 2 cotes, c'est ici le ratio entre la cote d'un groupe de blessés (ex : deux-roues motorisé sans tiers) et la côte du groupe de blessés de référence (pour le type d'usagers, le groupe de blessés de référence est « automobilistes avec tiers »)

Ainsi, les blessés en deux-roues motorisés dans un accident sans tiers ont une cote de 7,22 d'être enregistrés dans Registre seulement contre 1 dans BAAC seulement, par rapport aux blessés automobilistes avec tiers. Autrement dit les blessés en deux roues motorisés sans tiers sont bien mieux enregistrés par rapport aux blessés automobilistes avec tiers, dans Registre seulement que dans BAAC seulement.

On peut se baser sur l'ampleur des odds-ratios pour avoir une idée de l'importance de tel ou tel facteur. Ainsi les cyclistes blessés sans tiers sont sans commune mesure (OR=83) bien mieux enregistrés que les automobilistes avec tiers dans Registre seulement que dans BAAC seulement. Les blessés en voiture, sans tiers sont mieux enregistrés que les blessés en voiture avec tiers (OR=2,6) dans le Registre exclusivement que dans les BAAC exclusivement. Comme le registre est quasi exhaustif, cela signifie a contrario que les cyclistes sans tiers sont bien moins enregistrés dans les BAAC seulement que les automobilistes blessés dans un accident avec tiers.

Concernant l'année, les cotes sont de 1,05 d'être enregistré dans « Registre seulement » contre 1 dans « BAAC seulement » pour une année donnée (dans la période 2006-2012) par rapport à la précédente, en zone gendarmerie. Comme le registre est quasi exhaustif, cela signifie a contrario que l'enregistrement des blessés se dégrade au fil du temps en zone gendarmerie.

(NB : l'odds-ratio de l'ordonnée à l'origine correspond au groupe de blessés de référence du modèle, à savoir les blessés automobilistes, dans un accident avec tiers, en zone Gendarmerie, classés hospitalisés, dans un accident corporel, sur voies communales, l'année 2006).

Souvent, les odds ratios peuvent être interprétés comme des risques relatifs. Or, les conditions pour pouvoir interpréter les odds ratios de cette façon sont que l'odds-ratio n'est pas trop grand (classiquement moins de 3), ET que l'évènement étudié est rare. Or, pour un blessé, le fait d'être enregistré par les forces de l'ordre ou le registre n'est pas rare (NB : on n'étudie pas ici la fréquence de survenue d'un accident, qui est rare, mais la fréquence d'enregistrement d'un blessé, « une fois que c'est un blessé »).

Les coefficients de correction issus du modèle sont eux faciles à interpréter et sont donnés ci-après.

### 3.4.5 Exemple de coefficients de correction estimés sur 2006-2012

Nous rappelons qu'il s'agit des coefficients correcteurs du sous-enregistrement et des biais associés, autrement dit, pas seulement du sous-enregistrement. Les coefficients étant liés à la combinaison de toutes les variables présentes dans le modèle, ils prennent 267 valeurs différentes. Nous présentons ci-dessous les coefficients de correction du sous-enregistrement des données des forces de l'ordre pour quelques profils de blessés : les blessés dans un accident en 2011, sur route nationale ou départementale, enregistrés par la force de l'ordre Police (tableau ci-dessous), puis par la Gendarmerie (tableau ci-après), en laissant varier le type d'usager, la présence / absence de tiers, la gravité du blessé (hospitalisé ou non), et la gravité de l'accident (corporel / mortel).

*Tableau 20 : coefficients de correction du sous-enregistrement des données des forces de l'ordre pour quelques profils de blessés (accidents sur routes nationales et départementales, en zone police, en 2011)*

| type d'usagers | tiers | type d'accident | gravité | coefficient de correction des BAAC |
|---|---|---|---|---|
| usagers de 2RM | avec tiers | accident corporel | blessés hospitalisés | 1,1 |
| usagers de 2RM | avec tiers | accident corporel | blessés non hospitalisés | 1,5 |
| usagers de 2RM | avec tiers | accident mortel | blessés hospitalisés | 1,0 |
| usagers de 2RM | avec tiers | accident mortel | blessés non hospitalisés | 1,1 |
| usagers de 2RM | sans tiers | accident corporel | blessés hospitalisés | 2,2 |
| usagers de 2RM | sans tiers | accident corporel | blessés non hospitalisés | 7,5 |
| usagers de 2RM | sans tiers | accident mortel | blessés hospitalisés | 1,4 |
| usagers de 2RM | sans tiers | accident mortel | blessés non hospitalisés | 1,8 |
| cyclistes | avec tiers | accident corporel | blessés hospitalisés | 1,2 |
| cyclistes | avec tiers | accident corporel | blessés non hospitalisés | 1,9 |
| cyclistes | avec tiers | accident mortel | blessés hospitalisés | 1,1 |
| cyclistes | avec tiers | accident mortel | blessés non hospitalisés | 1,1 |
| cyclistes | sans tiers | accident corporel | blessés hospitalisés | 17,1 |
| cyclistes | sans tiers | accident corporel | blessés non hospitalisés | 91,8 |
| cyclistes | sans tiers | accident mortel | blessés hospitalisés | 7,2 |
| cyclistes | sans tiers | accident mortel | blessés non hospitalisés | 12,6 |
| automobilistes | avec tiers | accident corporel | blessés hospitalisés | 1,2 |
| automobilistes | avec tiers | accident corporel | blessés non hospitalisés | 2,0 |
| automobilistes | avec tiers | accident mortel | blessés hospitalisés | 1,1 |
| automobilistes | avec tiers | accident mortel | blessés non hospitalisés | 1,1 |
| automobilistes | sans tiers | accident corporel | blessés hospitalisés | 1,4 |
| automobilistes | sans tiers | accident corporel | blessés non hospitalisés | 3,2 |
| automobilistes | sans tiers | accident mortel | blessés hospitalisés | 1,1 |
| automobilistes | sans tiers | accident mortel | blessés non hospitalisés | 1,3 |
| piétons | avec tiers | accident corporel | blessés hospitalisés | 1,1 |
| piétons | avec tiers | accident corporel | blessés non hospitalisés | 1,7 |
| piétons | avec tiers | accident mortel | blessés hospitalisés | 1,0 |
| piétons | avec tiers | accident mortel | blessés non hospitalisés | 1,1 |
| autres | avec tiers | accident corporel | blessés hospitalisés | 1,1 |
| autres | avec tiers | accident corporel | blessés non hospitalisés | 1,3 |
| autres | avec tiers | accident mortel | blessés hospitalisés | 1,0 |
| autres | avec tiers | accident mortel | blessés non hospitalisés | 1,0 |
| autres | sans tiers | accident corporel | blessés hospitalisés | 1,4 |
| autres | sans tiers | accident corporel | blessés non hospitalisés | 3,3 |
| autres | sans tiers | accident mortel | blessés hospitalisés | 1,2 |
| autres | sans tiers | accident mortel | blessés non hospitalisés | 1,3 |

Ce tableau se lit de la manière suivante : pour les blessés en deux-roues motorisé, dans un accident sans tiers, dans un accident corporel, et classés blessés non hospitalisés par la police, il faut multiplier leur effectif par 7,5 afin d'obtenir l'effectif réel de ces blessés.

On note globalement des coefficients correcteurs plus grands pour les blessés non hospitalisés qu'hospitalisés, que les coefficients correcteurs des blessés dans un accident corporels sont plus élevés que ceux dans un accident mortel, que les blessés dans un accident sans tiers ont des coefficients correcteurs plus élevés pour ceux dans un accident avec tiers. Cela correspond au moindre enregistrement des blessés les moins graves, dans les accidents les moins graves, et dans les accidents sans tiers plutôt qu'avec tiers. En termes d'usagers, globalement, les piétons sont les moins à corriger ; viennent ensuite les automobilistes, puis les usagers « autres », puis les usagers de deux-roues motorisé et en dernier les cyclistes.

Le tableau ci-après présente les coefficients correcteurs en zone gendarmerie, en 2011.

*Tableau 21 : coefficients de correction du sous-enregistrement des données des forces de l'ordre pour quelques profils de blessés (accidents sur routes nationales et départementales, en zone gendarmerie, en 2011)*

| type d'usagers | tiers | type d'accident | gravité | coefficient de correction des BAAC |
|---|---|---|---|---|
| usagers de 2RM | avec tiers | accident corporel | blessés hospitalisés | 1,2 |
| usagers de 2RM | avec tiers | accident corporel | blessés non hospitalisés | 5,2 |
| usagers de 2RM | avec tiers | accident mortel | blessés hospitalisés | 1,0 |
| usagers de 2RM | avec tiers | accident mortel | blessés non hospitalisés | 1,2 |
| usagers de 2RM | sans tiers | accident corporel | blessés hospitalisés | 3,9 |
| usagers de 2RM | sans tiers | accident corporel | blessés non hospitalisés | 54,1 |
| usagers de 2RM | sans tiers | accident mortel | blessés hospitalisés | 1,5 |
| usagers de 2RM | sans tiers | accident mortel | blessés non hospitalisés | 3,6 |
| cyclistes | avec tiers | accident corporel | blessés hospitalisés | 1,4 |
| cyclistes | avec tiers | accident corporel | blessés non hospitalisés | 8,3 |
| cyclistes | avec tiers | accident mortel | blessés hospitalisés | 1,1 |
| cyclistes | avec tiers | accident mortel | blessés non hospitalisés | 1,4 |
| cyclistes | sans tiers | accident corporel | blessés hospitalisés | 40,8 |
| cyclistes | sans tiers | accident corporel | blessés non hospitalisés | 741,2 |
| cyclistes | sans tiers | accident mortel | blessés hospitalisés | 8,0 |
| cyclistes | sans tiers | accident mortel | blessés non hospitalisés | 37,5 |
| automobilistes | avec tiers | accident corporel | blessés hospitalisés | 1,5 |
| automobilistes | avec tiers | accident corporel | blessés non hospitalisés | 9,4 |
| automobilistes | avec tiers | accident mortel | blessés hospitalisés | 1,1 |
| automobilistes | avec tiers | accident mortel | blessés non hospitalisés | 1,4 |
| automobilistes | sans tiers | accident corporel | blessés hospitalisés | 2,0 |
| automobilistes | sans tiers | accident corporel | blessés non hospitalisés | 18,8 |
| automobilistes | sans tiers | accident mortel | blessés hospitalisés | 1,2 |
| automobilistes | sans tiers | accident mortel | blessés non hospitalisés | 1,9 |
| piétons | avec tiers | accident corporel | blessés hospitalisés | 1,3 |
| piétons | avec tiers | accident corporel | blessés non hospitalisés | 6,4 |
| piétons | avec tiers | accident mortel | blessés hospitalisés | 1,1 |
| piétons | avec tiers | accident mortel | blessés non hospitalisés | 1,3 |
| autres | avec tiers | accident corporel | blessés hospitalisés | 1,1 |
| autres | avec tiers | accident corporel | blessés non hospitalisés | 3,7 |
| autres | avec tiers | accident mortel | blessés hospitalisés | 1,0 |
| autres | avec tiers | accident mortel | blessés non hospitalisés | 1,1 |
| autres | sans tiers | accident corporel | blessés hospitalisés | 2,0 |
| autres | sans tiers | accident corporel | blessés non hospitalisés | 19,7 |
| autres | sans tiers | accident mortel | blessés hospitalisés | 1,2 |
| autres | sans tiers | accident mortel | blessés non hospitalisés | 1,9 |

Globalement, les coefficients correcteurs du sous-enregistrement et de ses biais sont d'amplitude bien plus grande qu'en zone police. On note des coefficients égaux à 1 (certains blessés dans un accident mortel), mais on obtient des coefficients égaux à 20, chez les automobilistes et usagers « autres » classés non hospitalisés, dans un accident corporel, sans tiers. Pour les usagers de deux-roues motorisés dans le même type d'accident, le coefficient correcteur atteint 54, et chez les cyclistes il atteint le record de 741. Ce chiffre énorme s'explique par la quasi absence d'enregistrement des cyclistes qui se blessent seuls dans les BAAC (et encore moins en zone Gendarmerie) : nous rappelons que les effectifs sont de 6 cyclistes blessés seuls par an en moyenne dans les BAAC rhodaniens versus 855 dans le Registre (sur la période 2006-2012).

## 3.5 Étape 3 : Projection du département du Rhône à la France entière

Ce qui nous intéresse, c'est l'extrapolation d'indicateurs de santé estimés au niveau local ou régional à une zone géographique plus large, régionale ou nationale. Les études publiées sur le sujet sont présentées en annexe. Une étude intéressante pour nous car similaire est l'estimation du nombre de toxicomanes consommateurs de cocaïne à partir du nombre observé d'héroïnomanes : la relation entre les deux est observée puis modélisées sur Londres, où les deux populations sont connues, puis appliquée au Royaume-Uni (Gossop, Strang et al. 1994). Dans le domaine de la traumatologie routière, une étude de 1991 est très proche de notre situation : le nombre de blessés de la route est donné par une source médicale dans 6 régions ou 6 hôpitaux (ce n'est pas précisé) et mis en relation, avec les données policières, semble –t-il en stratifiant sur le type d'usager et la gravité : des taux de correction moyens sont estimés sur ces 6 agrégats et appliqués aux données nationales policières du Royaume-Uni.

Actuellement, la Commission Européenne demande à ses Etats membres de fournir en 2015, le nombre de blessés de la route selon le critère médical MAIS3+ : certains pays européens vont estimer ces chiffres en appliquant aussi un facteur correctif aux données policières (d'autres pays vont se baser uniquement sur les données nationales hospitalières).

### 3.5.1 Application des coefficients correcteurs aux données nationales des forces de l'ordre

Ce que l'on projette ici, du département du Rhône à l'échelle du territoire France métropolitaine, ce n'est pas l'accidentalité observée sur le Rhône (éventuellement en fonction de diverses caractéristiques) mais les pratiques d'enregistrement des blessés des forces de l'ordre, du Rhône vers la France.

La méthode mise en œuvre est analogue à une standardisation indirecte (terminologie épidémiologique) d'un taux d'incidence, de cancer par exemple (Estève, Benhamou et al. 1994), effectuée classiquement sur âge et sexe. Ici il s'agit de la standardisation indirecte, de coefficients de correction de sous-enregistrement (au lieu d'incidences), sur les facteurs de sous-enregistrement (gravité de l'usager, gravité de l'accident, type d'usager, tiers, réseau et force de l'ordre) au lieu des facteurs de confusion âge et sexe.

L'estimation nationale du nombre de blessés est donnée par :

$$E_F = \sum_{j=1}^{J} C_j \times O_{F,j} \text{ , où :}$$

j indice les strates définies par la combinaison des facteurs de biais d'enregistrement, à la place des strates d'âge et de sexe,
$C_j$ sont les coefficients de correction du sous-enregistrement, de référence (Rhône) au lieu des taux d'incidences de référence, dans la strate j,
$O_{F,j}$ sont les effectifs observés des blessés selon les forces de l'ordre, de la population étudiée (F=France entière), au lieu des personnes-années de la population étudiée.

En écrivant l'étape intermédiaire :

$$E_F = \sum_{j=1}^{J} E_{F,j} = \sum_{j=1}^{J} C_j \times O_{F,j} \text{ ,}$$

nous pouvons utiliser la terminologie de la théorie des sondages (Ardilly 1994) ; voir annexes. Cela montre qu'il y a d'une part une post-stratification sur les facteurs du sous-enregistrement :

$$E_F = \sum_{j=1}^{J} E_{F,j} \text{ .}$$

Cela montre, qu'il y a d'autre part, dans chaque strate j, un redressement par le ratio :

$$E_{F,j} = C_j \times O_{F,j}.$$

Les dénombrements des forces de l'ordre ($O_{F,j}$) sont redressés par un coefficient ($C_j$) afin d'obtenir de meilleurs estimations, et ce coefficient est estimé sur un échantillon, réduit ici, à un département.

Cette formulation signifie, que pour une catégorie donnée, par exemple les cyclistes hospitalisés dans un accident corporel avec tiers sur une route départementale ou nationale en zone gendarmerie, nous considérons que le taux de sous-enregistrement modélisé sur le département du Rhône est une bonne estimation du taux de sous-enregistrement ailleurs en France. L'hypothèse sous-jacente est celle d'une homogénéité des pratiques des forces de l'ordre d'enregistrement des victimes sur l'ensemble de la France, à l'intérieur des sous-groupes définis par la combinaison des facteurs de sous-enregistrement.

Les strates sont définies par la combinaison des facteurs de biais de sélection : la gravité de l'accident (corporel/mortel), la gravité du blessé (hospitalisé ou non), le type d'usager, l'implication de tiers (oui/non), le type de réseau et la force de l'ordre. En ce qui concerne le type de force de l'ordre, nous assimilons la police aux frontières et la préfecture de police de Paris à la force de l'ordre « police ».

Le graphique ci-dessous reprend le principe de la projection. La méthode de capture-recapture permet d'estimer directement l'effectif total du sur-ensemble des blessés de la route (au niveau rhodanien) et indirectement d'estimer, entre les effectifs des BAAC rhodanien et ce sur-ensemble rhodanien, des coefficients de correction du sous-enregistrement des BAAC.

Ces coefficients de correction sont ensuite appliqués aux BAAC nationaux, ou à l'estimation qui en est faite par redressement des blessés des PV corporels au 1/20ème, et il sera ainsi obtenu des résultats représentatifs de la réalité de l'accidentalité au niveau national.

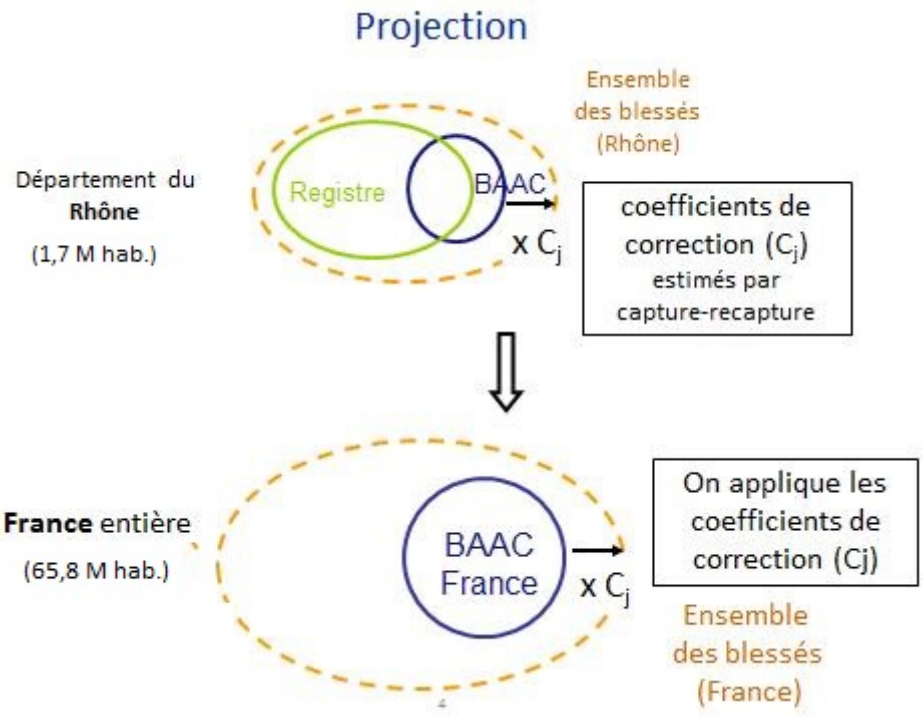


*Figure 8: projection, du Rhône à la France entière*

# 4 AFFECTATION DES COEFFICIENTS ET UTILISATION DES PONDERATIONS

## 4.1 Affectation des coefficients correcteurs des biais d'enregistrements aux données

Les coefficients correcteurs sont indicés par j ou m. Nous rappelons que cette lettre représente les différentes caractéristiques, qui influent sur la probabilité d'enregistrement dans les BAAC : type de force de l'ordre (CRS / gendarmerie / police), gravité de l'accident (corporel ou mortel), gravité du blessé (hospitalisé ou non), type d'usager (automobiliste / usager de 2RM / cycliste / piéton / autre), type d'accident (avec ou sans tiers), type de réseau (autoroute, route nationale, départementale, communale ou autre), et année.

Pour affecter ces coefficients aux blessés, nous utilisons l'information des PV codée par les experts de VOIESUR, qui peut être différente dans quelques cas, de celle des BAAC. Si l'information est manquante dans le PV VOIESUR, nous utilisons alors l'information fournie par les BAAC. Cela arrive surtout pour la variable gravité, où notamment la durée d'hospitalisation peut être inconnue.

Les coefficients correcteurs sont définis au niveau de l'unité statistique blessé. La question qui se pose est : quel coefficient correcteur utiliser lorsque l'unité statistique étudiée est différente ? Il faut garder en tête qu'il s'agit de corriger en fonction inverse de la probabilité d'enregistrement dans les BAAC.

Si l'unité statistique étudiée est la lésion, on lui affecte le coefficient correcteur du blessé. Il se peut que toutes les lésions n'aient pas la même probabilité d'être enregistrées, mais c'est justement ce que la tâche 4.X va étudier en comparant les descriptions lésionnelles de VOIESUR et celles du Registre du Rhône sur les blessés communs.

Si l'unité statistique étudiée est l'accident (corporel), on lui affecte le minimum des coefficients correcteurs des blessés impliqués dans l'accident. C'est-à-dire qu'on pense que la probabilité d'enregistrement de l'accident dépend de la probabilité d'enregistrement la plus grande des blessés de l'accident.

Si l'unité statistique étudiée est le véhicule, on lui affecte le coefficient correcteur de l'accident, et non le minimum des coefficients correcteurs des blessés du véhicule en question. Car sinon, on pourrait aboutir à deux coefficients correcteurs différents pour les 2 véhicules d'un même accident, et cela nous semble peu plausible. Pour les accidents à plus de 2 véhicules, il est possible que la probabilité d'enregistrement décroisse avec le nombre de véhicules mais il ne nous est pas possible de la déterminer.

Cas particuliers :
Pour les indemnes, conducteurs ou passagers, on leur affecte le coefficient correcteur de l'accident (c'est-à-dire le minimum des coefficients correcteur des blessés de l'accident).

(Nous rappelons que pour les accidents mortels, les tués et l'accident lui –même sont considérés correctement enregistrés ; ils n'ont pas de coefficient correcteur (ou bien, égal à 1).

## 4.2 Utilisation des différents poids dans les analyses statistiques (tâche 4)

Le tableau ci-après rappelle quels poids, au sens large, c'est-à-dire incluant aussi coefficient de redressement ou coefficient correcteur, sont à utiliser.

*Tableau 22 : utilisation des différents poids (sondage, redressement sous-enregistrement et ses biais)*

| | Poids de sondage (pk) | Contrainte sur effectifs totaux (constante, incluse dans Rl) | Coefficient de redressement des PV au 1/20ème vers les BAAC (Rl) | Coefficient correcteur (Cj) du sous-enregistrement et de ses biais |
|---|---|---|---|---|
| S'applique à : | Toutes les observations si on analyse ensemble accidents mortels (pk=1) et accidents corporels (pk= 20) | Blessés dans les accidents corporels au 1/20ème<br>Blessés dans les accidents mortels | Tous les PV corporels au 1/20ème, quelle que soit l'unité statistique associée (blessé, accident, lésion,..) | Tous les blessés (d'un PV mortel ou corporel), et quelle que soit l'unité statistique associée |
| Sont exclus : | | - tués<br>-blessés Rhône pris dans leur ensemble | - PV Rhône pris dans leur ensemble<br>- PV mortels | tués |

En pratique pour utiliser les trois types de poids à la fois, il faut créer une variable poids, qui est le produit des différents poids :

poids= $p_k$ * $R_l$ *$C_j$ si les 3 types de poids sont considérés pertinents

Dans les analyses statistiques, il faut alors :

- déclarer cette variable en tant que poids (instruction WEIGHT dans le logiciel SAS),
- utiliser les procédures statistiques permettant de prendre en compte un plan de sondage (procédures SURVEYXXXXX dans SAS), et
- préciser si besoin les aspects suivants : la population d'étude doit être considérée comme infinie (par défaut dans SURVEYFREQ et SURVEYLOGISTIC dans SAS) et les poids doivent être normalisés (par défaut dans SURVEYFREQ et SURVEYLOGISTIC dans SAS), c'est-à-dire que leur somme doit être rendue égale à la taille de l'échantillon.

Sans ces précautions, les variances, et donc les intervalles de confiance ne seront pas correctement estimés, et les estimations ponctuelles dans des modèles de régression pourront aussi être incorrectes.

# 5 DISCUSSION

## 5.1 Effet Registre du Rhône sur les forces de l'ordre du Rhône?

Dans les analyses en cours des PV de VOIESUR, il est apparu que les PV du Rhône étaient un peu mieux renseignés que les PV France entière. Ainsi, la proportion de PV comprenant un bilan lésionnel est de 67% dans les PV 1/20ème France entière versus 83% dans le Rhône. De même, la

vitesse initiale lors du choc a pu être estimée dans 36% des PV 1/20ème France entière vs 49% des PV du Rhône, et pour la vitesse au choc : 11.5% versus 18.8%. Il semble donc qu'il y ait un effet du registre dans le sens d'un meilleur contenu des PV. Est-ce qu'il y aurait aussi se traduire par un meilleur enregistrement des blessés dans le Rhône qu'ailleurs ?
Si tel était le cas cela signifierait que les coefficients de correction du sous-enregistrement, estimés sur le Rhône sous-estiment les coefficients correcteurs qu'il faudrait appliquer ailleurs. Cela signifie que les effectifs totaux de blessés estimés sous-estiment la réalité.
Il est à noter que cela va dans le même sens, c'est-à-dire une sous-estimation des effectifs de blessés, que la dépendance positive entre l'enregistrement des blessés dans les deux sources, BAAC et Registre, au niveau du Rhône.

## 5.2 Validité des modèles et de l'ensemble de la procédure d'extrapolation

D'après Tilling et Sterne (Tilling and Sterne 1999) les tests habituels d'adéquation d'un modèle aux données ne sont pas utilisables car ce sont des tests d'adéquation aux données observées, alors que l'approche capture-recapture repose sur l'existence de données non observées (les victimes non recensées).
En revanche, afin d'évaluer la validité de notre procédure, nous comparons nos résultats avec des estimations extérieures quand elles existaient.

### 5.2.1 Elément externe de validation : nombre de traumatisés médullaires

Une première comparaison porte sur le nombre incident de victimes avec traumatisme médullaire. D'après la caisse nationale de l'assurance maladie, il y a eu 2343 nouveaux cas de paraplégies (IIS 4 et 5) en moyenne annuelle sur 1996-2004 (il s'agit des nouvelles prises en charge en affection de longue durée (CNAM 2007)). Par ailleurs, des chercheurs (Saillant, Pascal-Moussellard et al. 2005) estiment, en appliquant les taux d'incidence d'autres pays industrialisés, États-Unis essentiellement, qu'il y a environ 2000 cas annuels de traumatismes médullaires. Il est aussi estimé que 60-70 % (ibid.) de ces traumatismes sont dus à des accidents de la voie publique, qui sont essentiellement des accidents de la circulation (les autres étant des accidents de piétons seuls, comptabilisés comme accidents de la vie courante et ayant peu de risques d'aboutir à une lésion médullaire). Cela donne environ 1200-1400 traumatisés médullaires suite à un accident de la circulation routière chaque année.

Par ailleurs, notre méthode d'extrapolation estime ce nombre à 1145 (IC à 95 % = 853-1438), moyenne annuelle sur 1996-2004. Ces effectifs sont très proches.

### 5.2.2 Elément externe de validation : l'Enquête Nationale Transport et Déplacements (ENTD)

Une deuxième validation est fournie par l'ENTD (Enquête Nationale Transports Déplacements) qui s'est déroulée en 2007-2008. Cette enquête se base sur 20200 personnes (âgées de 6 ans et plus). Il leur est demandé les accidents corporels des 5 dernières années, de la façon suivante : « (prénom) a-t-il/elle été victime d'un accident de la circulation entraînant des dommages corporels au cours des 5 dernières années ? (accidents ayant provoqué une blessure donnant lieu à un acte médical : médecin, hôpital…) ». Les réponses à cette question avaient permis d'estimer (Haddak M (sous la coordination de), Bouaoun et al. 2013) le nombre de blessés de la route, toutes gravités, par année, France entière (tableau ci-après).

*Tableau 23: estimation du nombre de blessés selon l'ENTD (2007-2008), personnes âgées de 6 ans et plus*

| année | Effectif estimé |
|---|---|
| 2003 | 369 303 |
| 2004 | 323 796 |
| 2005 | 399 581 |
| 2006 | 469 789 |
| 2007 | 416 544 |

On note une certaine fluctuation d'une année à l'autre, avec sans doute un biais de mémoire : les personnes se rappellent moins bien des accidents les plus éloignés dans le temps, notamment les accidents les moins graves, mais on a globalement un ordre de grandeur autour de 400 000.

Par extrapolation, on trouvait 400 236 blessés toutes gravités pour 2004 auxquels il faut retrancher les blessés de moins de 6 ans ( 5676 blessés de 0-4 ans et grosso modo 1/5ème des 8560 blessés de 5-9 ans) soit 392248. Pour rappel, les blessés toutes gravités selon les BAAC sont au nombre de 108 727 en 2004, et en enlevant les blessés de moins de 6 ans, on obtient 106 565.

Les premières estimations (provisoires) obtenues France entière sur 2006-2012, du nombre de blessés toutes gravités sont du même ordre de grandeur que celles sur 1996-2004 : on obtient 378500 blessés toutes gravités en 2006 et 399 000 en 2007, auxquels on retranche les blessés de moins de 6 ans, ce qui nous donne 371 259 blessés en 2006 et 392 069 en 2007. C'est donc un peu en dessous des estimations de l'ENTD, mais dans le même ordre de grandeur autour de 400 000.

### 5.2.3 Elément externe de validation : ratio blessés toutes gravités /tués

Un dernier élément de validation vient du ratio blessés toutes gravités / tués. Ce ratio ne devrait pas être très différent entre pays européens semblables, ou plutôt devrait être du même ordre de grandeur. Dans le graphique ci-après on donne les ratios de quelques pays de l'union européenne pour 2009. Le ratio blessés toutes gravités/tués en France, non corrigé, vaut 14, en dessous de la moyenne européenne à 34, et aussi en dessous de l'Espagne et de l'Italie. En corrigeant du sous-enregistrement des BAAC, le ratio français atteint alors 70, ce qui le situe au même niveau que l'Allemagne et le Royaume-Uni, sur leurs données non corrigées.

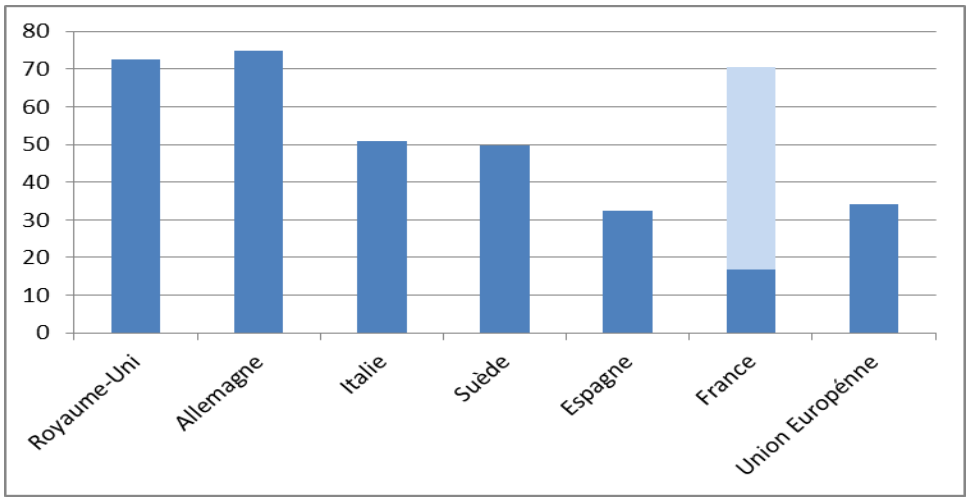


*Figure 9 : ratios blessés toutes gravités / tués, 2009*

Ces trois éléments extérieurs viennent confirmer l'ordre de grandeur des estimations obtenues par la méthode d'extrapolation.

# 6 CONCLUSION

Nous rappelons que cette tâche 3 de VOIESUR comprend un redressement des blessés des PV corporels au 1/20ème afin d'être représentatifs de l'ensemble des BAAC corporels, suivi d'une extrapolation, qui corrige les données de type BAAC du sous-enregistrement et surtout des biais associés.

Le redressement est réalisé par post-stratification sur la variable force de l'ordre, qui parait la plus importante dans la différence observée entre PV au 1/20ème de VOIESUR et BAAC.

La correction du sous-enregistrement et surtout des biais associés se fait par méthode de capture-recapture, et se traduit par l'estimation de coefficients correcteurs. Ils dépendent de différentes caractéristiques, celles liées à la probabilité d'enregistrement dans les BAAC : type de force de l'ordre (CRS / gendarmerie / police), gravité de l'accident (corporel ou mortel), gravité du blessé (hospitalisé ou non), type d'usager (automobiliste / usager de 2RM / cycliste / piéton / autre), type d'accident (avec ou sans tiers), type de réseau (autoroute, route nationale, départementale, communale ou autre), et année.

Pour les analyses de la tâche 4 de VOIESUR , il faut créer une variable poids, produit multiplicateur entre le poids de sondage ( 1 pour les accidents mortels, 20 pour les accidents corporels), le coefficient de redressement et le coefficient correcteur du sous-enregistrement et de ses biais. Cette variable doit être déclarée comme poids dans les analyses statistiques. Ainsi les résultats seront représentatifs de l'accidentalité de la route sur la France entière.

Deux rappels importants pour toute inférence utilisant des données pondérées :

- Il est indispensable que les estimations de variances, et les tests d'hypothèses correspondants prennent correctement en compte les pondérations.
- Ceci ne dispense pas d'un examen critique des estimations obtenues sur des éléments fortement pondérés mais rarement observés.

## 7 RÉFÉRENCES


Abeni, D. D., G. Brancato, et al. (1994). "Capture-recapture to estimate the size of the population with human immunodeficiency virus type 1 infection." Epidemiology **5**(4): 410-414.

Alho, J. (1990). "Logistic regression in capture-recapture models." Biometrics **46**: 623-635.

Amoros, E. (2007). Les blessés par accidents de la route : estimation de leur nombre et de leur gravité lésionnelle, France, 1996-2004 ; modélisation à partir d'un registre médical (Rhône) et des données policières (France) thèse de doctorat, Université Lyon 1.

Amoros, E., J. L. Martin, et al. (2008). "Actual incidences of road casualties, and their injury severity, modelled from police and hospital data, France." European Journal of Public Health **18**: 360-365.

Amoros, E., J. L. Martin, et al. (2006). "Under-reporting of road crash casualties in France." Accident Analysis and Prevention **38**(4): 627-635.

Amoros, E., J. L. Martin, et al. (2008). "Estimation de la morbidité routière, France, 1996-2004." Bulletin Epidemiologique Hebdomadaire **19**: 157-160.

Aptel, I., L. R. Salmi, et al. (1999). "Road accident statistics: discrepancies between police and hospital data in a French island." Accident Analysis and Prevention **31**(1-2): 101-108.

Ardilly, P. (1994). Les techniques de sondage. Paris, Editions Technip.

Bernillon, P., L. Lievre, et al. (2000). "Record-linkage between two anonymous databases for a capture-recapture estimation of underreporting of AIDS cases: France 1990-1993." International Journal of Epidemiology **29**(1): 168-174.

Bloor, M., A. Leyland, et al. (1991). "Estimating hidden populations: a new method of calculating the prevalence of drug-injecting and non-injecting female street prostitution." Bristish Journal of Addiction **86**(11): 1477-1483.

Brenner, H. and I. Schmidtmann (1996). "Determinants of homonym and synonym rates of record linkage in disease registration." Methods of Information in Medicine **35**(1): 19-24.

Brenner, H., C. Stegmaier, et al. (1994). "Estimating completeness of cancer registration in Saarland/Germany with capture-recapture methods." European Journal of Cancer **30A**(11): 1659-1663.

Brenner, H., C. Stegmaier, et al. (1995). "Estimating completeness of cancer registration: an empirical evaluation of the two source capture-recapture approach in Germany." Journal of Epidemiology and Community Health **49**(4): 426-430.

Chao, A., P. K. Tsay, et al. (2001). "The applications of capture-recapture models to epidemiological data." Statistics in Medecine **20**(20): 3123-3157.

Clark, D. E. (2004). "Practical introduction to record linkage for injury research." Injury Prevention **10**(3): 186-191.

CNAM. (2007). "Affection Longue Durée - incidence." Retrieved 3 juillet 2007, from http://www.ameli.fr/l-assurance-maladie/statistiques-et-publications/donnees-statistiques/affection-de-longue-duree-ald/index.php.

Cormack, R. M. (1989). "Loglinear models for capture-recapture." Biometrics **45**: 395-413.

Crocetti, E., G. Miccinesi, et al. (2001). "An application of the two-source capture-recapture method to estimate the completeness of the Tuscany Cancer Registry, Italy." European Journal of Cancer Prevention **10**(5): 417-423.

Deville, J. and C. Sarndal (1992). "Calibration Estimators in Survey Sampling." Journal of the American Statistical Association **87**(418): 376-382.

Dhillon, P. K., A. S. Lightstone, et al. (2001). "Assessment of hospital and police ascertainment of automobile versus childhood pedestrian and bicyclist collisions." Accident Analysis and Prevention **33**(4): 529-537.

Estève, J., E. Benhamou, et al. (1994). Descriptive epidemiology. Lyon, International agency for research on cancer.

Fienberg, S. E. (1992). "Bibliography on capture-recapture modelling with applications to census undercount adjustments." Survey methodology **18**: 143-154.

Fisher, N., S. W. Turner, et al. (1994). "Estimating numbers of homeless and homeless mentally ill people in north east Westminster by using capture-recapture analysis." Bristish Medical Journal **308**(6920): 27-30.

Foster, P. J. and G. J. Johnson (2001). "Glaucoma in China: how big is the problem?" British Journal of Ophthalmology **85**(11): 1277-1282.

Frey, C. M., E. J. Feuer, et al. (1994). "Projection of incidence rates to a larger population using ecologic variables." Statistics in Medecine **13**(17): 1755-1770.

Frischer, M., M. Bloor, et al. (1991). "A new method of estimating prevalence of injecting drug use in an urban population: results from a Scottish city." International Journal of Epidemiology **20**(4): 997-1000.

Frischer, M., M. Hickman, et al. (2001). "A comparison of different methods for estimating the prevalence of problematic drug misuse in Great Britain." Addiction **96**(10): 1465-1476.

Frischer, M., A. Leyland, et al. (1993). "Estimating the population prevalence of injection drug use and infection with human immunodeficiency virus among injection drug users in Glasgow, Scotland." American Journal of Epidemiology **138**(3): 170-181.

Gallay, A., A. Nardone, et al. (2002). "La méthode capture-recapture appliquée à l'épidémiologie : principes, limites et applications." Revue d'Epidémiologie et de Santé Publique **50**(2): 219-232.

Garton, M. J., M. I. Abdalla, et al. (1996). "Estimating the point accuracy of population registers using capture-recapture methods in Scotland." Journal of Epidemiology and Community Health **50**(1): 99-103.

Gemmell, I., T. Millar, et al. (2004). "Capture-recapture estimates of problem drug use and the use of simulation based confidence intervals in a stratified analysis." Journal of Epidemiology and Community Health **58**(9): 758-765.

Gossop, M., J. Strang, et al. (1994). "A ratio estimation method for determining the prevalence of cocaine use." Bristish Journal of Psychiatry **164**(5): 676-679.

Haddak M (sous la coordination de), L. Bouaoun, et al. (2013). Rapport ISOMERR-Ménages n°3 : Mesure du risque d'accident de la route à l'aide d'indicateurs d'exposition au risque. R. d. c. P. (GO2). Bron (69), Ifsttar-Umrestte. **3:** 103 p.

Hay, G., N. McKeganey, et al. (1999). Methodological guidelines to estimate the prevalence of problem drug use on the local level. Lisbon, European Monitoring Centre for Drugs and Drug Addiction**:** 1-76.

Hilsenbeck, S. G., C. Kurucz, et al. (1992). "Estimation of completeness and adjustment of age-specific and age-standardized incidence rates." Biometrics **48**(4): 1249-1262.

Hook, E. B. and R. R. Regal (1995). "Capture-recapture methods in epidemiology: methods and limitations." Epidemiologic Reviews **17**(2): 243-264.

Hook, E. B. and R. R. Regal (1999). "Recommendations for presentation and evaluation of capture-recapture estimates in epidemiology." Journal of Clinical Epidemiology **52**(10): 917-926; discussion 929-933.

Hook, E. B. and R. R. Regal (2000). "Accuracy of alternative approaches to capture-recapture estimates of disease frequency: internal validity analysis of data from five sources." American Journal of Epidemiology **152**(8): 771-779.

Hook, E. B. and R. R. Regal (2000). "On the need for a 16th and 17th recommendations for capture-recapture analysis." Journal of Clinical Epidemiology **53**(12): 1275-1277.

Howe, G. R. (1998). "Use of computerized record linkage in cohort studies." Epidemiologic Reviews **20**(1): 112-121.

IWGDMF (1995). "International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation I: History and theoretical development." American Journal of Epidemiology **142**(10): 1047-1058.

IWGDMF (1995). "International Working Group for Disease Monitoring and Forecasting. Capture-recapture and multiple-record systems estimation II: Applications in human diseases. ." American Journal of Epidemiology **142**(10): 1059-1068.

James, H. F. (1991). "Under-reporting of road traffic accidents." Traffic Engineering and Control **32**: 573-583.

Jarvis, S. N., P. J. Lowe, et al. (2000). "Children are not goldfish - mark/recapture techniques and their application to injury data." Injury Prevention **6**: 46-50.

Jensen, O. M., J. Esteve, et al. (1990). "Cancer in the European Community and its member states." European Journal of Cancer **26**(11-12): 1167-1256.

Johnson, R. L., B. A. Gabella, et al. (1997). "Evaluating sources of traumatic spinal cord injury surveillance data in Colorado." American Journal of Epidemiology **146**(3): 266-272.

Johri, M., E. Kaplan, et al. (1999). "New approaches to HIV surveillance: means and ends. Summary report of conference held at Yale University, 21-22 May 1998,

by the Law, Policy and Ethics Core, Center for Interdisciplinary Research on AIDS, Yale University." AIDS Public Policy Journal **14**(4): 136-146.

Jones, I. E., R. Cannan, et al. (2000). "Distal forearm fractures in New Zealand children: annual rates in a geographically defined area." New Zealand Medicine Journal **113**(1120): 443-445.

LaPorte, R. E., S. R. Dearwater, et al. (1995). "Efficiency and accuracy of disease monitoring systems: application of capture-recapture methods to injury monitoring." American Journal of Epidemiology **142**(10): 1069-1077.

Madigan, D. and J. York (1997). "Bayesian methods for estimation of the size of a closed population." Biometrika **84**(1): 19-31.

Maxwell, J. C. (2000). "Methods for estimating the number of "hard-core" drug users." Substance Use and Misuse **35**(3): 399-420.

Menegoz, F., R. J. Black, et al. (1997). "Cancer incidence and mortality in France in 1975-95." European Journal of Cancer Prevention **6**(5): 442-466.

Meuleners, L. B., A. H. Lee, et al. (2006). "Estimating crashes involving heavy vehicles in Western Australia, 1999-2000: A capture-recapture method." Accident Analysis and Prevention **38**(1): 170-174.

Morrison, A. and D. H. Stone (2000). "Capture-recapture: a useful methodological tool for counting traffic related injuries?" Injury Prevention **6**(4): 299-304.

Nachbaur, C., Z. Uhry, et al. (2004). Estimation du taux d'incidence annuel. d'accidents de la vie courante, en France, en 2001. Journées scientifiques de l'InVS., Paris.

Neugebauer, R. and J. Wittes (1994). "Voluntary and involuntary capture-recapture samples--problems in the estimation of hidden and elusive populations." American Journal of Public Health **84**(7): 1068-1069.

Newcombe, H. B. (1988). Handbook of record linkage: methods for health and statistical studies, administration, and business. Oxford, Oxford university press.

Peabody, J. W., B. Schau, et al. (2005). "COPD: A prevalence estimation model." Respirology **10**(5): 594-602.

Razzak, J. A. and S. P. Luby (1998). "Estimating deaths and injuries due to road traffic accidents in Karachi, Pakistan, through the capture-recapture method." International Journal of Epidemiology **27**: 866-870.

Roberts, I. and R. Scragg (1994). "Application of capture-recapture methodology to estimate the completeness of child injury surveillance." Journal of Paediatrics and Child Health **30**(6): 513-514.

Robles, S., L. Marrett, et al. (1988). "An application of capture-recapture methods to the estimation of completeness of cancer registration." Journal of Clinical Epidemiology **41**(5): 495-501.

Rosenman, K. D., A. Kalush, et al. (2006). "How much work-related injury and illness is missed by the current national surveillance system?" Journal of Occupational and Environmental Medicine **48**(4): 357-365.

Saillant, G., H. Pascal-Moussellard, et al. (2005). "Les lésions traumatiques de la moelle épinière : épidémiologie et prise en charge hospitalière." Bulletin de l'Académie Nationale de Médecine **189**(6): 1095-1106.

Sautory, O. (1993). La macro CALMAR, redressement d'un échantillon par calages sur marges. Documents de travail. D. d. S. D. e. Sociales. Paris, INSEE. **F 9310:** 51 p.

Schouten, L. J., H. Straatman, et al. (1994). "The capture-recapture method for estimation of cancer registry completeness: a useful tool?" International Journal of Epidemiology **23**(6): 1111-1116.

Sherman, G. (1981). Cancer incidence in Canada: completeness and ecological correlations, University of Toronto.

Stutts, J. C., J. E. Williamson, et al. (1990). "Bicycle accidents and injuries: A pilot study comparing hospital- and police-reported data." Accident Analysis and Prevention **22**(1): 67-78.

Takala, J. (1999). "Global estimates of fatal occupational accidents." Epidemiology **10**(5): 640-646.

Tercero, F. and R. Andersson (2004). "Measuring transport injuries in a developing country: an application of the capture-recapture method." Accident Analysis and Prevention **36**(1): 13-20.

Tilling, K. and J. A. Sterne (1999). "Capture-recapture models including covariate effects." American Journal of Epidemiology **149**(4): 392-400.

# ANNEXES

## Annexe 1 : Méthodes de redressement, dans le cadre de la théorie des sondages

Nous rappelons ici ce qu'est le redressement dans le cadre de la théorie des sondages (Ardilly 1994). En général, on cherche à redresser un échantillon lorsque l'échantillon obtenu par plan de sondage n'est pas assurément représentatif de la population cible. Le redressement se définit comme une amélioration de l'estimation (dans une étude avec un plan de sondage) en utilisant de l'information auxiliaire.
L'idée est d'utiliser une information auxiliaire X corrélée à la variable d'intérêt Y. On doit avoir suffisamment d'informations sur X pour pouvoir redresser l'échantillon (avec un système de poids) de telle sorte qu'il soit représentatif de la répartition de X dans la population cible. La corrélation de X avec Y conduit à ce que l'échantillon ainsi redressé donne une meilleure estimation de Y.

Trois méthodes de redressement sont définies :
- redressement par post-stratification
- redressement par le ratio
- redressement par régression

1) Redressement par post-stratification :

Il s'agit d'un redressement sur une ou plusieurs variables auxiliaires, catégorielles ou rendues catégorielles. Celles-ci forment des post-strates, par opposition aux variables de stratification utilisées dans le plan de sondage. L'information auxiliaire se résume à la connaissance de la répartition dans la population générale de ces post-strates (et non de la répartition dans l'échantillon).
Au lieu de l'estimateur brut, on prend pour estimateur une moyenne pondérée des estimateurs par strate, où le poids est l'effectif relatif de la strate dans la population.

Ex : estimer le revenu moyen de la population française. Le revenu est corrélé avec l'âge ; on redresse donc l'estimation en utilisant la répartition des tranches d'âge dans la population française, et non celle de l'échantillon, qui peut par hasard ne pas être représentative de la population.

$$\hat{\bar{Y}}_R = \sum_{h=1}^{H} w_h \times \bar{y}_h = \sum_{h=1}^{H} \frac{N_h}{N} \times \bar{y}_h ,$$

avec h=identifiant des post-strates, $w_h$ le nouveau poids

Post-stratification en fonction de plusieurs variables auxiliaires :
Si on dispose des effectifs pour toutes les strates définies par le croisement de toutes les modalités des variables auxiliaires : même méthode que celle au-dessus.

Si on ne dispose des informations que sur les marges, c'est-à-dire pour une variable auxiliaire à la fois, on utilise une procédure de calage sur les marges (Deville and Sarndal 1992; Sautory 1993), qui ressemble à la post-stratification sur une variable, mais avec des itérations pour tenir compte de plusieurs variables auxiliaires.

2) Redressement par le ratio ou par le quotient

Pour améliorer l'estimation, on utilise une variable auxiliaire (quantitative) qui a une relation de proportionnalité avec la variable d'intérêt.

on suppose : $Y_i = R \times X_i + U_i$, avec X variable auxiliaire, et Ui résidus de somme nulle
et donc l'estimateur redressé s'écrit :

$\hat{\overline{Y}} = \overline{y} \times \frac{\overline{X}}{\overline{x}}$, avec $R \approx \frac{\overline{y}}{\overline{x}}$, estimé sur l'échantillon

pondération (des observations) :

$$w_i(s) = \frac{1}{n} \times \frac{\overline{X}}{\overline{x}}.$$

Cela signifie que l'on redresse par un coefficient correcteur ; ce coefficient est égal au ratio entre la valeur moyenne de l'information auxiliaire dans la population cible et celle dans l'échantillon.

3) Redressement par régression

Pour améliorer l'estimation, on utilise une variable auxiliaire (quantitative) qui est en relation linéaire avec la variable d'intérêt.

on suppose : $Y_i = a + b \times X_i + U_i$, avec X variable auxiliaire, et Ui résidus de somme nulle

et donc l'estimateur redressé s'écrit : $\hat{\overline{Y}} = \overline{y} + \hat{b} \times \left(\overline{X} - \overline{x}\right)$.

On redresse les observations récoltées au niveau de l'échantillon par une régression.

Il est facile d'étendre le redressement par régression à plusieurs variables auxiliaires.

## Annexe 2 : Synthèse bibliographique sur la projection, du régional au national, en épidémiologie

Ce qui nous intéresse, c'est l'extrapolation d'indicateurs de santé estimés au niveau local ou régional à une zone géographique plus large, régionale ou nationale. Nous n'avons pas considéré les études avec un plan de sondage sur tout un territoire ; pour celles-ci, il s'agit essentiellement de faire de l'inférence, en tenant compte du plan de sondage.

Par indicateurs de santé, nous entendons en particulier l'incidence, c'est-à-dire le nombre de nouveaux cas sur une année en général ; le taux d'incidence est le nombre de nouveaux cas rapportés à la population exposée, qui peut être la population générale. Par exemple, le taux d'incidence de cancers, est le nombre de personnes nouvellement diagnostiquées comme atteintes d'un cancer, sur une année, rapportée à la population générale. Pour les accidents de la route, le taux d'incidence de blessés de la route est le nombre de blessés de la route, sur une année donnée, rapporté à la population générale.

Nous avons trouvé des publications en épidémiologie des addictions (drogues), du cancer, des traumatismes… et nous avons pu classer ces études, selon les méthodes implicites ou explicites utilisées. Nous présentons ci-dessus les méthodes que nous avons dégagées, avec les exemples correspondants.

**Extrapolation brute**
Le taux d'incidence observé sur une région est utilisé comme taux d'incidence du pays. Précisément, l'incidence observée sur une région est appliquée à (l'effectif de) la population nationale, et on obtient ainsi le nombre de cas incidents au niveau national. Cela correspond à la règle de trois : on a trois effectifs, qui sont le nombre de cas incidents sur une région, la population de la région, la population nationale, et par « règle de trois » on obtient le nombre de cas incidents de la population nationale.

Exemple :
Estimation du nombre d'accidents professionnels mortels dans le monde, à partir d'incidences observées dans certains pays, puis appliquée aux autres pays, selon la similarité entre pays (Takala 1999)

Hypothèse sous-jacente :
Le taux d'incidence observé dans une région, est supposé être le même que celui de la région d'application (ou du territoire tout entier).

**Extrapolation en ajustant sur des facteurs de confusion**

a) Extrapolation en ajustant sur catégories d'âge et de sexe, ou autres, définis au niveau individuel
Principe :
L'incidence est observée sur une région, par strates de sexe et d'âge, et appliquée, par strates de sexe et d'âge, à la population générale.

Exemples :
- estimation du nombre de personnes avec fractures distales de l'avant-bras sur l'ensemble de la Nouvelle-Zélande, à partir de l'incidence observée, par sexe et groupe d'âge, sur une ville majeure, Dunedin, 111000 habitants (Jones, Cannan et al. 2000)

- estimation de la prévalence de glaucome en chine : application de la prévalence observée en Mongolie, à la Chine rurale, et application de la prévalence observée à Singapour, à la Chine urbaine, par sexe et groupe d'âge (Foster and Johnson 2001).

Cela revient à pondérer les taux d'incidence par les effectifs par strate d'âge et de sexe, i.e. à de la post-stratification (cf. annexe) sur âge et sexe ; il s'agit aussi de standardisation indirecte, selon la terminologie en épidémiologie, où le nombre attendus de cas incidents de la population étudiée est estimé en appliquant le taux d'incidence, observé sur une population de référence, à la distribution d'âge et de sexe de la population étudiée.

Hypothèse sous-jacente :
Le taux d'incidence dans une région, par strate d'âge et de sexe, est le même, à l'intérieur de chaque strate, que celui de la région d'application (ou du territoire tout entier).

b) Extrapolation en ajustant sur des facteurs de confusion, définis à un niveau agrégé
Principe :
Une relation entre l'incidence d'une maladie et des caractéristiques régionales est observée et modélisée, et on applique ensuite cette relation à la population générale.

Exemples :
- estimation de l'incidence de cancer sur tous les Etats-Unis à partir de la relation entre l'incidence de cancer et des variables socio-économiques et ethniques ; relation établie pour certains Etats, puis appliquée aux autres Etats (Frey, Feuer et al. 1994)
- estimation de la prévalence de broncho-pneumopathie chronique obstructive (COPD), en fonction de sa relation avec la structure démographique, la proportion de fumeurs, la proportion d'urbains, le développement économique, la pollution environnementale… : relation observée et modélisée dans certains pays, puis appliquée à d'autres pays (Peabody, Schau et al. 2005).

Détail de la méthode (avec variables définies à un niveau agrégé):
Etape 1 : obtention, sur plusieurs régions, de l'incidence observée et de variables corrélées agrégées et observées ;
Etape 2 : modélisation de l'incidence (observée) en fonction des variables régionales observées ; la modélisation est de type modèle linéaire généralisé ;
Etape 3 : application du modèle obtenu aux autres régions, où les mêmes variables régionales sont observées, afin d'obtenir une estimation du nombre des incidents sur ces régions ; en sommant sur les régions, on obtient une estimation nationale.
Cela correspond à la méthode de redressement par régression (cf. annexe), où les États ou pays constituent les individus de l'échantillon.

Hypothèse sous-jacente :
La relation entre le taux d'incidence et certaines caractéristiques macroscopiques (démographiques, économiques,...) est supposée être la même dans la région d'observation et dans la région d'application (ou dans le territoire tout entier).

**Extrapolation à partir d'une morbidité corrélée ou de la mortalité associée**

a) Extrapolation en fonction d'un indicateur similaire, et d'éventuels facteurs de confusion, définis à un niveau individuel

Principe :

L'incidence d'intérêt est modélisée par une relation proportionnelle à l'incidence d'une morbidité associée mieux connue, ou au taux de mortalité associé à la morbidité d'intérêt, sur une (des) région(s) où les deux sont observées. Extrapolation aux autres régions où seule l'incidence de la morbidité associée ou le taux de mortalité est observé. (Il y a éventuellement stratification sur âge et sexe).

Exemples :

- estimation de la prévalence de toxicomanes consommateurs de cocaïne à partir de la prévalence observée d'héroïnomanes : relation observée et modélisée sur Londres, puis appliquée au Royaume-Uni (Gossop, Strang et al. 1994)
- estimation de l'incidence des cancers (par site) en fonction du taux de mortalité de ces cancers, par strate d'âge et de sexe ; pour la France, relation entre incidence et mortalité observée et modélisée dans les départements avec registres de cancer, puis appliquée à toute la France (Jensen, Esteve et al. 1990).

En épidémiologie des traumatismes :

- estimation du nombre de blessés de la route traités dans les services d'urgences, proportionnellement au nombre de blessés de la route hospitalisés ; relation observée sur deux groupes d'hôpitaux de l'Etat de Caroline du Nord (au nombre de 10 et 15 hôpitaux) et appliquée à tout l'Etat (Stutts, Williamson et al. 1990).
- estimation du nombre de blessés de la vie courante traités dans les services d'urgences, proportionnellement au nombre d'hospitalisés pour traumatisme ; relation observée sur trois hôpitaux du réseau EPAC, Enquête permanente sur les accidents de la vie courante (Bordeaux, Annecy, Béthune) et appliquée à la France entière (Nachbaur, Uhry et al. 2004).

Détail de la méthode utilisant l'incidence d'une morbidité associée :

Etape 1 : obtention, sur une (ou des) région(s), de l'incidence d'intérêt et de l'incidence d'une morbidité similaire et corrélée ;

Etape 2 : modélisation de l'incidence d'intérêt (observée) en fonction de l'incidence (observée) de cette morbidité similaire ; la modélisation est de type proportionnelle ; et éventuellement par strate d'âge et sexe.

Etape 3 : application du modèle obtenu aux autres régions où l'incidence de cette morbidité associée est observée, afin d'obtenir une estimation du nombre des cas incidents pour la morbidité d'intérêt sur ces régions.

Cela correspond à un redressement par le ratio (cf. annexe), et éventuellement par post-stratification.

Hypothèse sous-jacente :

La relation observée et modélisée (entre l'incidence d'intérêt et une morbidité associée) sur la région d'observation est supposée être la même dans les régions d'application.

b) Extrapolation en fonction d'un indicateur similaire et de caractéristiques macroscopiques

Principe :

L'incidence d'intérêt est modélisée par une relation à partir de l'incidence d'une morbidité associée mieux connue ou de la mortalité associée, et de caractéristiques générales sur une (des) région(s) où toutes sont observées. La relation modélisée est appliquée aux autres régions où seule l'incidence de

la morbidité associée et les caractéristiques générales sont observées. La modélisation est de type modèle linéaire généralisé.

Exemples :
- estimation de l'incidence des cancers (par site) en fonction du taux de mortalité de ces cancers, de l'âge, et de l'année, en France (avec étape supplémentaire de lissage des données de mortalité) ; relation entre incidence et mortalité observée et modélisée dans les départements avec registres de cancer et appliquée à toute la France (Menegoz, Black et al. 1997).
- estimation du nombre de toxicomanes en fonction du nombre de décès liés à l'usage de drogue, du nombre de toxicomanes admis en traitement médical, du nombre de toxicomanes verbalisés par la police, de la quantité de drogue saisie, et du nombre de séropositifs par injection de drogue ; relation observée sur quatre régions (parmi une quinzaine du Royaume-Uni) et appliquée à l'ensemble du Royaume-Uni (Frischer, Hickman et al. 2001).
Cela correspond au redressement par régression (cf. annexe), où les régions ou départements ou villes constituent les individus de l' « échantillon ».

Hypothèse sous-jacente :
La relation observée et modélisée sur la région d'observation est supposée être la même dans les régions d'application.

Enfin, nous avons rencontré un exemple un peu différent : l'indicateur similaire est exactement le même mais évalué par une autre source : estimation du nombre de victimes en fonction du nombre de victimes recensées par la police, par application de taux de correction moyens (et semble-t-il en stratifiant sur le type d'usager, et sur la gravité) : relation observée sur 6 études (sans doute 6 régions ou 6 hôpitaux), et appliquée à la Grande-Bretagne (James 1991).

Cet exemple correspond à notre situation. L'article indiquait que ces 6 études n'étaient pas forcement représentatives, et que le résultat donnait juste une indication de l'ampleur réelle de la morbidité routière. On peut analyser cette approche comme étant un redressement par post-stratification, et à l'intérieur de chaque strate, d'un redressement par le ratio.