Recueil du Symposium 2018 de Statistique Canada
Combiner pour vaincre : innovations dans l'utilisation de multiples sources de données

Comprendre les effets du couplage d'enregistrements sur l'estimation du total lors de la combinaison d'une source de mégadonnées avec un échantillon probabiliste

Benjamin Williams1[1]

Résumé

Le National Marine Fisheries Service (NMFS) estime le nombre total de poissons capturés par les pêcheurs à la ligne dans les eaux américaines. NMFS arrive à ce nombre en multipliant l'estimation de l' « effort » (nombre de voyages) par les prises par unité d' « effort » ou nombre de poissons capturés par voyage. Les données sur l' « effort » sont recueillies au moyen d'une enquête postale. Les données sur les poissons capturés par voyage (PUE) sont recueillies au moyen d'enquêtes en personne en interceptant les pêcheurs au quai après des voyages de pêche. Le NMFS explore la possibilité de remplacer l'enquête sur l' « effort » par l'auto-déclaration volontaire. Les pêcheurs à la ligne signalent les détails de leurs voyages au moyen d'un appareil électronique et demeurent admissibles à l'enquête par interception au quai. Les estimateurs proposés du total utilisent les rapports dans le cadre des modèles de capture-recapture (Liu et coll., 2017). Pour obtenir une estimation valide, les données des rapports et des voyages interceptés nécessitent d'être couplées. L'appariement exact est difficile à atteindre en pratique en raison d'erreurs non dues à l'échantillonnage. Dans ce présent document, nous élaborons un algorithme de couplage d'enregistrements pour coupler les voyages, et examiner l'effet qu'a le choix du point de découpage sur l'estimation finale.

Mots-clés : Échantillonnage; échantillonnage non probabiliste; couplage d'enregistrements; erreurs d'appariement.

1. Introduction

1.1 Contexte

Aux États-Unis, les pêches sont classées en deux catégories : commerciales et récréatives. Les pêches commerciales sont tenues de déclarer leur nombre total de prises. Ce n'est pas le cas pour les pêcheurs à la ligne. Dans certaines régions, le nombre total de poissons capturés par les pêcheurs à la ligne peut dépasser le nombre total de prises de la pêche commerciale (Conseil national de recherches, 2017). Estimer ce nombre total de prises est donc d'une extrême importance.

Connaître le total des prises de poissons est essentiel pour établir la longueur de la saison de pêche, les limites de prises, les prises biologiques acceptables, et la limite de prises annuelle pour les diverses espèces de poisson (Conseil national de recherches, 2017). Les estimations sont des intrants aux modèles d'abondance de poisson. Les extrants des modèles sont utilisés pour la gestion des pêches afin de maintenir la stabilité des niveaux des populations, prévenir la surpêche, et lutter contre les effets des catastrophes naturelles comme les déversements de pétrole, qui peuvent avoir une incidence négative sur les populations de poissons (Tarnecki et Patterson, 2015).

Le National Marine Fisheries Service (NMFS) de la National Oceanic and Atmospheric Administration (NOAA) estime le nombre total de prises de poissons par les pêcheurs à la ligne par l'entremise du Marine Recreational Information Program (MRIP). Le produit de deux estimations est utilisé pour estimer les prises totales : effort (*E*), qui est le nombre de voyages effectués, et les prises par unité d'effort (*PUE*), qui est la moyenne des prises de chaque espèce par voyage de pêcheur à la ligne. Le MRIP estime l'effort et les prises par unité d'effort par l'entremise de

[1]Benjamin Williams, Southern Methodist University, 3225, avenue Daniel, Dallas, TX, É.-U., 75205 (benjamin@smu.edu)

deux enquêtes. Chaque enquête est un échantillon probabiliste ce qui signifie que chaque unité d'échantillonnage a une probabilité connue d'être sélectionnée dans l'échantillon.

Le premier échantillon est un échantillon obtenu par interception à quai, appelé Access Point Angler Intercept Survey (APAIS). L'APAIS est un échantillon des quais publics. L'unité primaire d'échantillonnage (UPE) est une combinaison d'emplacements de quai public et une heure sur un jour précis dans un délai de deux mois appelé une vague. L'unité secondaire d'échantillonnage (USE) est un voyage effectué par un pêcheur à la ligne. Un plan d'échantillonnage de probabilité proportionnelle à la taille (PPT) est utilisé pour sélectionner les intervieweurs de l'UPE. Les intervieweurs visitent chaque UPE sélectionné pour interviewer tous les pêcheurs à la ligne qui reviennent d'un voyage de pêche. Les intervieweurs notent des statistiques comme le nombre de poissons capturés par espèces et par pêcheur à la ligne, le nombre de pêcheurs à bord du bateau, le nombre de poissons remis à l'eau par espèce, etc. À partir de l'échantillon APAIS, les prises par unité d'effort(*PUE)*, pour chaque espèce, sont estimées.

Le deuxième échantillon est l'enquête d'effort de pêche (EEP). L'EEP est une enquête envoyée par la poste aux résidents vivant dans les États qui bordent le golfe du Mexique. L'EEP est une enquête dont la base de sondage est fondée sur les adresses et est un échantillon de tous les pêcheurs à la ligne, sur le National Saltwater Registry, qui vivent dans un état directement adjacent d'une zone de pêche en eau salée (comme le golfe du Mexique). Les répondants de l'EEP sont invités à fournir leurs efforts (le nombre de fois où ils sont allés à la pêche) rétrospectivement pour la dernière vague (deux mois) pour jusqu'à cinq membres du ménage (Conseil national de recherches, 2017). Un ajustement pour les pêcheurs à la ligne qui ne vivent pas dans les États échantillonnés est effectué au moyen d'une proportion estimative des pêcheurs résidant dans les États de l'échantillon de l'interception à quai. Le nombre total des prises de poissons par les pêcheurs à la ligne est enfin estimé en multipliant les *PUE* et l'*E* pour chaque espèce.

Cette méthodologie actuelle est le résultat de plusieurs années de mises à jour méticuleuses et intentionnelles. Le Conseil national de recherches (CNR) a examiné le MRIP et formulé des recommandations à deux reprises, une fois en 2006 puis de nouveau en 2016. Il y a encore des problèmes dans la procédure d'estimation du MRIP dont plusieurs sont décrits dans le plus récent examen du CNR.

La majorité des problèmes dans les techniques d'estimation concernent l'EEP. La première lacune est due aux erreurs de mesure. L'EEP a lieu à la fin d'une vague et demande aux répondants de se rappeler le nombre de voyages de pêche effectués au cours de cette vague. Cela veut dire qu'ils doivent se rappeler d'événements qui ont eu lieu il y a plus de deux mois. Les répondants peuvent avoir de la difficulté à répondre à cette question avec précision, surtout s'ils sont des pêcheurs à la ligne enthousiastes. Un autre problème est lié à l'efficacité de la production de l'estimation. Le NMFS rapporte qu'il faut environ 45 jours après chaque vague pour obtenir les estimations finales du nombre total de prises. Cela s'explique par le temps nécessaire pour effectuer la livraison et obtenir suffisamment de réponses pour une estimation valide. Une estimation plus rapide pourrait permettre l'établissement à temps opportun des limites de pêche.

Enfin, parce que les *PUE* sont estimées à partir de l'APAIS, elle est produite en utilisant uniquement les voyages de retour à sites publics seulement. Une hypothèse implicite à la multiplication des *PUE* et de l'*E* est que le nombre de *PUE* reste le même pour les voyages de retour vers des quais publics et privés. Cela peut être une hypothèse raisonnable, mais il se pourrait également que les voyages privés attrapent plus ou moins de poissons par pêcheur à la ligne et par voyage. Ceci est une autre raison de passer de l'EEP à la déclaration électronique, parce que l'autodéclaration permet de reporter le nombre total des prises pour les voyages de retour vers des quais privés.

Lors de l'examen du CNR de 2016, le conseil a proposé plusieurs recommandations pour remédier à ces lacunes et améliorer l'estimation. La recommandation, qui motive cette recherche, a suggéré une évaluation des méthodes de collecte de données électroniques pour possiblement remplacer l'EEP (Conseil national de recherches, 2017). Le CNR a noté que ces méthodes de déclaration électroniques pourraient permettre une estimation en temps quasi réel. Dans certaines régions des États-Unis, les institutions responsables de la gestion des pêches ont commencé à expérimenter ce genre de techniques. Dans le golfe du Mexique, le NMFS expérimente actuellement dans plusieurs États en demandant aux capitaines d'autodéclarer leurs voyages avec un appareil électronique.

Le NMFS a collaboré avec une firme de recherche privée (appelée CLS) pour cette expérience. Les capitaines de bateaux récréatifs peuvent se porter volontaires pour participer. CLS fournit un appareil électronique aux bénévoles, ce qui leur permet de déclarer eux-mêmes les données ayant trait à la démographie et à la pêche pour leurs voyages

de pêche récréatifs. Parce que l'auto-déclaration se produit sur un appareil électronique, les données sont disponibles pour estimation en temps quasi réel. Un capitaine peut être sélectionné dans l'échantillon d'interception et faire rapport de son voyage avec l'appareil électronique, ce qui veut dire que le voyage peut être présent dans les deux échantillons. Le but de l'expérience est de remplacer l'EEPdont la base de sondage est fondée sur les adresses, par l'échantillon volontaire des capitaines qui s'auto-déclarent. Toutefois, cet échantillon volontaire est un échantillon non probabiliste, et donc la méthode actuelle d'estimation n'est pas valide. Ces deux échantillons sont suffisants pour estimer le nombre total de prises parce qu'ils constituent une forme de modèle de capture-recapture.

1.2 Méthodologie actuelle

Les méthodes de capture-recapture sont de puissants moyens d'estimer un total dans certains scénarios. Dans un exemple classique, supposons qu'un chercheur désire connaître le nombre total de poissons (N) dans le coin de pêche local. Lors du premier voyage de pêche, cette personne attrape n_1 poissons. Ces poissons reçoivent une étiquette pour être identifiés plus tard. Le lendemain, elle retourne au coin de pêche et attrape n_2 poissons. Dans cette deuxième prise, supposons que m poissons avaient déjà été capturés le premier jour, et étaient identifiables par leur étiquette. Si la proportion de poissons étiquetés dans le deuxième échantillon est à peu près égale à la proportion de poissons étiquetés dans la population de poissons, alors $E\left(\frac{n_1}{N}\right) \approx \frac{m}{n_2}$. Cela conduit à l'estimateur Lincoln-Peterson du total (Cren, 1956) :

$$\widehat{N} = \frac{n_1 n_2}{m} \tag{1.1}$$

$\widehat{N}$ est l'estimateur du maximum de vraisemblance dans le cadre du modèle hypergéométrique, ce qui suppose que l'échantillon de recapture est un échantillon aléatoire simple (EAS). Cette hypothèse EAS ne s'étend pas nécessairement à l'échantillon initial.

Dans l'expérience du NMFS, l'échantillon auto-déclaré est analogue à la portion « capture » d'un programme de capture-recapture, tandis que l'échantillon d'interception à quai est la composante de « recapture ». Parce que l'APAIS est un échantillon probabiliste, les estimateurs peuvent être semblables à l'estimateur Lincoln-Peterson de l'équation 1.1. Cependant, certaines des unités de la capture, les voyages auto-déclarés, qui retournent vers des quais privés, n'ont aucune chance d'être incluses dans l'échantillon de recapture. Ainsi, utiliser (1.1) exige que le taux de déclaration pour les voyages de retour vers des quais publics et privés soient équivalents. Liu et coll. (2017) ont enquêté sur cette situation au Texas et ont produit des estimateurs cohérents sur les prises totales en adaptant $\widehat{N}$ à partir de (1.1).

Premièrement, définir l'univers d'intérêt comme étant les N voyages de pêche récréative dans le golfe du Mexique. Définir les prises pour certaines espèces dans le i^e voyage comme y_i $(i = 1, \dots, N)$. L'objectif consiste à estimer $t_y = \sum_{i=1}^{N} y_i$. Dans les données auto-déclarées, les prises signalées pour le i^evoyage sont notées par y_i^*. Si le i^e voyage n'est pas signalé, y_i^* est défini comme étant 0. y_i^* diffère de y_i parce qu'il y a une erreur de mesure causée par des incohérences entre le rapport du capitaine et les données de l'intervieweur.

Puisqu'il est utile de considérer les déclarants comme des domaines, notons l'échantillon probabiliste (APAIS) par s_2 et l'échantillon non probabiliste (auto-déclarations électroniques) par d_1. Il y a n_2 voyages échantillonnés dans s_2 et n_1 voyages signalés dans d_1. Nous examinons un estimateur dans ce scénario recommandé par Liu et coll. (2017). Cet estimateur est appelé $\hat{t}_{y2}$ et est un estimateur de ratio multivarié (Olkin, 1958). Il est représenté par cette équation :

$$\hat{t}_{y2} = t_{y^*} + \frac{n_1}{\hat{n}_1}\left(\hat{t}_y - \hat{t}_{y^*}\right) \tag{1.2}$$

où $t_{y^*} = \sum_{i \in d_1} y_i^*$, $\hat{n}_1 = \sum_{i \in s_2} r_i w_i$, $\hat{t}_y = \sum_{i \in s_2} w_i y_i$, $\hat{t}_{y^*} = \sum_{i \in s_2} r_i y_i^*$. Ici, $r_i = 1$ si la i^e unité d'échantillonnage a été signalée, et est 0 sinon; w_i est le poids d'échantillonnage, défini comme l'inverse de la probabilité d'échantillonnage pour la i^e unité. Liu et coll. (2017) souligne que $\hat{t}_{y2}$ est un sous-dénombrement moyen ajusté des prises ajouté au total des prises déclarées. Les estimations de $\hat{n}_1$et $\hat{t}_{y^*}$ nécessitent de coupler les voyages qui ont été à la fois signalés et interceptés.

2. Couplage d'enregistrements

2.1 Erreurs d'appariement

L'exactitude du processus de couplage peut avoir une grande incidence sur la qualité des estimateurs des prises. Par conséquent, il est important de s'assurer que les couplages sont aussi exacts que possible. À l'origine, nous croyions que le numéro d'identification du bateau, la date et l'heure du voyage, et le lieu de son retour fourniraient un lien unique qui produirait une correspondance parfaite, puisque les renseignements sur ces variables sont enregistrés à la fois dans les données de l'enquête d'interception et les données auto-déclarées. Cependant, ce n'était pas le cas. Pour les rapports, le capitaine déclare une partie des renseignements (p. ex., le nombre de passagers) et une autre partie provient directement de l'appareil électronique (p. ex., l'emplacement). Ces deux sources sont sujettes à des erreurs, mais en raison de différentes causes. Les renseignements sur l'emplacement dans les rapports sont en fait une série d'emplacements GPS que l'appareil signale à des intervalles de 15 minutes. L'enquête d'interception, quant à elle, nous fournit un numéro d'identification et le nom d'une marina ou d'un autre emplacement sur le cadre à partir duquel les UPE de l'interception sont choisies. Ainsi, le même voyage signalé et intercepté n'aura pas des emplacements GPS identiques, mais devrait plutôt être simplement proche. D'autres erreurs non dues à l'échantillonnage qui rendent le couplage difficile comprennent les erreurs liées à l'appareil et le fait que plusieurs capitaines soumettent leurs rapports longtemps après la fin de leur voyage de pêche.

Initialement, nous avons effectué une opération de couplage à la main, avec une règle qui filtrait les voyages qui étaient « proches » selon le numéro d'identification du bateau, la date et l'heure de chaque voyage, et l'emplacement de fin d'un voyage. Toutefois, nous n'avons repéré qu'un petit nombre de voyages correspondants en utilisant cette approche. En raison du grand nombre de rapports et de notre connaissance du nombre d'appareils électroniques déployés, il semblait peu probable que le nombre de voyages obtenu était aussi petit que ce que nous avons constaté. Nous avons décidé de relâcher les critères requis pour identifier une correspondance, ou de nous fier à d'autres nombreuses variables déclarées dans les deux fichiers. Pour mener à bien une telle méthode, nous avions besoin d'une manière raisonnée d'aller de l'avant. Cela nous a conduits à la littérature sur le couplage des enregistrements.

2.2 Couplage d'enregistrements

Le couplage d'enregistrements est un processus visant à fusionner deux fichiers de données ou plus en fonction de variables présentes dans les deux sources de données. Lorsqu'il n'y a pas d'identificateur unique commun aux fichiers de données, le couplage d'enregistrements est utilisé pour les coupler. Pour coupler les voyages, nous suivons les techniques de couplage d'enregistrements de Fellegi et Sunter (1959) et de Bell et coll. (1994).

Fellegi et Sunter (1959) formalise le couplage d'enregistrements et nous allons maintenant résumer leur travail. Les deux fichiers sont dénotés par A et B. La série de toutes les paires ordonnées possibles de couplages entre les deux fichiers est dénotée par:
$A \times B \{(\mathrm{a}, \mathrm{b}) : \mathrm{a} \in \mathrm{A}, \mathrm{b} \in \mathrm{B}\}$. Cette série est l'union de la série de paires appariées et non appariées; c.-à-d., $[\mathrm{M} = \{(\mathrm{a}, \mathrm{b}) : \mathrm{a} = \mathrm{b}, \mathrm{a} \in \mathrm{A}, \mathrm{b} \in \mathrm{B}\}]$ et $\mathrm{U} = \{(\mathrm{a}, \mathrm{b}) : \mathrm{a} \neq \mathrm{b}, \mathrm{a} \in \mathrm{A}, \mathrm{b} \in \mathrm{B}\}]$. Leur objectif est de produire une règle de couplage qui classe chaque membre de $A \times B$ dans l'une des trois catégories possibles : soit il est déclaré un appariement (A_1), un appariement possible (A_2), ou un non-appariement (A_3). La règle de couplage est fondée sur la comparaison entre a et b sur une série de variables relatives au couplage et de déterminer si elles s'accordent. Le résultat de cette comparaison est un score. Les points de découpage pour le score définissent la règle de couplage. Les liens avec les scores élevés sont affectés à A_1 tandis que ceux ayant des scores bas sont affectés à A_3 (Fellegi et Sunter, 1959).

Bell, Keesey, et Richards (1994), définissent qu'un *appariement* survient lorsque deux enregistrements (un de chaque ensemble de données) désignent la même unité. Un *couplage* survient lorsqu'on détermine que deux enregistrements désignent la même entité dans les deux fichiers (par l'entremise de quelque procédure d'appariement). Ils disent que deux enregistrements s'*accordent* lorsque les enregistrements présentent les mêmes valeurs sur les variables relatives au couplage, mais ne sont pas nécessairement un appariement. Ils ont créé un score de ce genre en se basant sur Fellegi et Sunter (1959), qui ajoute ou soustrait du poids en fonction du nombre de concordances entre les valeurs des variables relatives au couplage.

Dénotons par x et y les deux valeurs enregistrées pour cette variable pour une paire (a, b). En supposant que les enregistrements qui ne constituent pas des appariements sont jumelés au hasard, la note pour la k^e variable relative au couplage peut être écrite comme suit :

$$S_k = log\big(P(y|x, M = 1)\big) - log\big(P(y)\big) \quad (2.1)$$

où M est l'indicateur d'un appariement. Le score de correspondance de 2,1 a une forme unique pour chacune des trois situations possibles : x et y *s'accordent*, sont *proches*, ou *ne s'accordent pas*. Les scores pour chaque situation, respectivement, sont les suivants :

$$S_k = -log\big(P(y)\big) \quad (2.2)$$

$$S_k = log\big(P(\text{x et y sont proches}|M = 1)\big) - log\big(P(\text{un enregistrement aléatoire est proche de x}|M = 0)\big) \quad (2.3)$$

$$S_k = log\big(P(\text{x et y ne s'accordent pas}|M = 1)\big). \quad (2.4)$$

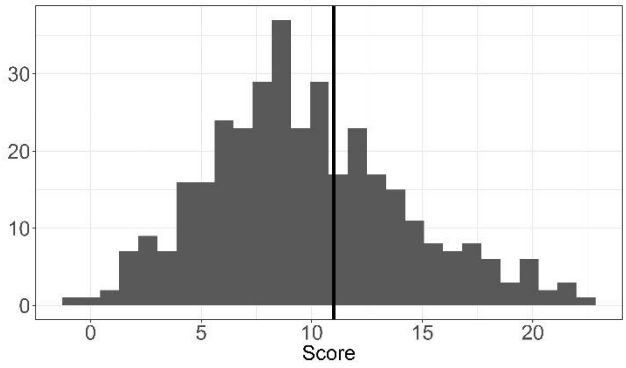
L'équation 2.2 suppose que la probabilité que x et y s'accordent pour un appariement est proche de 1, l'équation 2.3 suppose que $x\ P(\text{x et y sont proches}|M = 1, x)$ sont indépendants, et l'équation 2.4 suppose que la valeur y ne s'accorde pas aléatoirement avec la valeur x pour une certaine proportion des correspondances (Bell, Keesey, et Richards, 1994). Ensuite, chaque pièce des équations 2.2 à 2.4 doit être une estimation. Le score dans 2.1 est estimé de façon empirique à partir de l'un des fichiers (dans notre cas, nous avons utilisé des données d'interception). Pour estimer la composante de 2.2 conditionnée sur les non-appariements, nous calculons de façon empirique à partir de $A \times B$, sous l'approximation que presque tous les couples sont des non-appariements. Pour estimer les morceaux de 2.3 et 2.4 de façon conditionnelle aux enregistrements constituant un appariement, on suppose qu'un sous-ensemble de deux fichiers de données qui s'accordent sur quelques variables relatives au couplage est essentiellement un ensemble d'appariements. La probabilité pour les variables omises relatives au couplage est estimée en utilisant cette série. Cette méthodologie est élaborée en suivant Bell, Keesey, et Richards (1994).

Nos deux fichiers sont le fichier des interceptions de pêcheurs à la ligne et le fichier des voyages signalés. Nous avons des données de deux années d'expérience du NMFS, 2016 et 2017. En 2016, 1 569 voyages ont été notés dans l'interception à quai et 6 514 voyages ont été signalés. En 2017, il y a eu 1 380 voyages échantillonnés dans l'interception à quai et 9 132 voyages auto-déclarés. La qualité des données, surtout pour le fichier des voyages auto-déclarés était plus pauvre en 2016 qu'en 2017, étant donné que l'expérience venait tout juste de commencer et que des lacunes devaient encore être résolues, y compris les problèmes avec les numéros d'identification des navires. NOAA a préparé et présenté le fichier d'interception lors de leur cycle normal de production des données. Les variables disponibles à partir des deux fichiers sont presque identiques, mais la méthode de collecte est différente.

Pour le couplage, nous avons utilisé le numéro d'identification du navire en tant que seule variable pour les pochettes. Nous nous attendons à ce que certaines correspondances ne s'accordent pas au moins légèrement en ce qui concerne la date de retour au quai après un voyage, alors la date n'était pas une variable pour les pochettes. Il y a de nombreuses options pour coupler les variables, y compris la date de son retour au quai, l'État dans lequel le voyage est revenu, le nombre de pêcheurs à la ligne, le nombre total de prises de poissons par tous les pêcheurs à la ligne, le nombre total d'espèces capturées, la latitude/longitude du quai ou de l'emplacement de retour, le nombre individuel de capture et de remise à l'eau pour plus de 40 espèces de poissons, et d'autres. Nous avons choisi le numéro d'identification du navire, le nombre total de prises, le nombre total de rejets, le nombre de pêcheurs à la ligne sur le voyage, la latitude/longitude de l'emplacement de retour enregistré par l'appareil électronique des auto-déclarants et l'emplacement du quai précisé dans l'échantillon à quai, la date à laquelle le voyage a eu lieu, le nombre d'espèces capturées et le nombre d'espèces remises à l'eau.

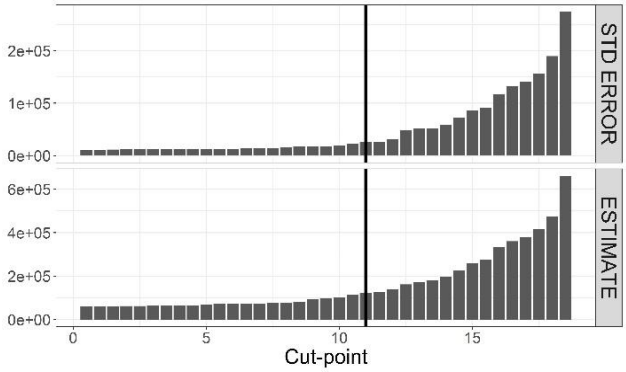
Après le couplage des enregistrements, un score de correspondance a été calculé pour chaque navire unique potentiellement auto-déclarant et la date/l'heure consignée dans s_2. Nous n'avons gardé que l'auto-déclaration avec le score le plus élevé. En cas d'égalité des deux voyages auto-déclarés, le couplage avec la plus petite distance entre les sites d'auto-déclaration et d'interception a été gardé. Si toute auto-déclaration était répétée dans les données, nous les avons écartés et pour ces voyages tirés de s_2, nous avons obtenu la correspondance la plus proche aux auto-déclarations. Après la procédure d'appariement, il y avait 591 voyages uniques avec un score pour les deux années.

Figure 1.2-1
Couplage des enregistrements des scores de 2017, point de découpage de 11 mis en évidence



L'étape suivante consistait à déterminer un score de découpage. Les paires d'enregistrements avec des scores inférieurs au point de découpage ne sont pas couplées, tandis que les paires d'enregistrements avec des scores supérieurs au point de découpage sont couplées. Pour 2017, nous avons choisi un point de découpage de 11 parce qu'il y a un creux au score de 11. Une distribution idéale du score est bimodale et penche vers la droite. Bien que la distribution de la Figure 2.2-1 soit légèrement asymétrique, un score de découpage n'est pas évident. La Figure 2.2-2 montre la valeur estimée du total pour l'espèce de poissons vivaneau rouge en 2017 ainsi que son erreur type en fonction du point de découpage. Le point de découpage a manifestement un effet.

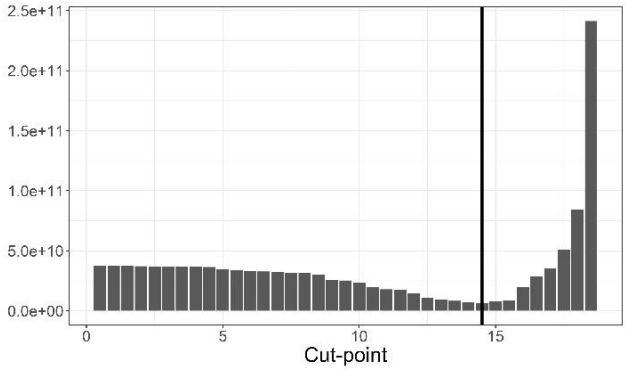
Figure 2.2-2
Estimation et erreur type de la récolte totale de vivaneau rouge (2017), point de découpage de 11 mis en évidence



Pour choisir le bon point de découpage, Fellegi et Sunter ont défini le point de découpage de façon à ce que les niveaux spécifiés des taux d'erreur pour les faux positifs et les faux négatifs soient atteints (Fellegi et Sunter, 1959). C'est certainement raisonnable, mais dans notre cas, nous ne pouvons déterminer avec certitude quels voyages constituent de vraies correspondances ou de vraies non-correspondances. Contrairement aux problèmes habituels de couplage des enregistrements dans lesquels des noms sont impliqués et un examen par le commis peut conduire à la bonne détermination du statut de correspondance, nous étudions deux voyages de pêche et comparons leurs variables numériques relatives au couplage. Il peut y avoir d'importantes erreurs de mesure dans ces variables, du moins de la part des voyages déclarés, ce qui ajoute à la difficulté de l'examen manuel

Une première idée pour déterminer le point de découpage est de comparer ces estimations aux estimations publiquement affichées sur le site Web de la NOAA pour avoir une idée de la partialité du biais de $\hat{t}_{y2}$. Même s'il s'agit d'une méthode insatisfaisante, en raison du possible biais dans les estimations de la NOAA elles-mêmes, il est toujours utile de comprendre les effets du point de découpage. En combinant cette idée de biais avec l'erreur type associée à $\hat{t}_{y2}$, nous pouvons obtenir un « pseudo-erreur quadratique moyen » de $\hat{t}_{y2}$ en tant que fonction du point de découpage (Figure 2.2-3).

Figure 2.2-3
« Pseudo-EQM » en tant que fonction du point de découpage (2017), point de découpage de 11 mis en évidence



Comme le montrent les figures 2.2-2 et 2.2-3, le choix initial d'un point de découpage de 11 choisi par simple inspection de la distribution du score (figure 2.2-1) ne minimise pas l'erreur type ou « pseudo-EQM » de $\hat{t}_{y2}$. Le choix d'un point de découpage mène à de nombreuses questions, notamment : comment peut-on déterminer un point de découpage qui aura une incidence sur les estimations de différentes espèces de poissons de différentes façons? Peut-être qu'un plus petit point de découpage donne de meilleures estimations pour le vivaneau rouge, mais pas pour la Gorette blanche. Nous hésitons à choisir un point de découpage fondé uniquement sur des estimations faites pour une seule espèce.

3. Travaux à venir

Ce travail est loin d'être terminé. Nous travaillons actuellement sur l'estimation des taux d'erreur pour les faux positifs et les faux négatifs afin d'essayer de choisir un point de découpage dans le même ordre d'idée de Fellegi et Sunter (1959). Nous avons également décrit un modèle pour les erreurs d'appariement afin d'introduire le hasard dans la procédure d'appariement. Cela nous permettra d'examiner les effets de diverses erreurs d'appariement sur le biais et la variance de $\hat{t}_{y2}$ et d'autres estimateurs proposés dans la littérature. Nous sommes également en train d'achever une étude de simulation pour examiner la nature complexe de l'échantillon d'interception. La simulation nous permettra d'étudier les effets de l'erreur d'appariement et du couplage d'enregistrements sur les estimations du total. Nous croyons que ce travail de mélange des échantillons a une valeur incroyable dans l'ère actuelle des méga-données, ainsi que dans l'exploitation du potentiel des échantillons non probabilistes.

Bibliographie

Bell, R.M., J. Keesey, et T. Richards (1994), « The Urge to Merge: Linking Vital Statistics Records and Medicaid Claims », *Medical Care*, 32, p. 1004 – 1018.

Cren, E. D. L. (1956). « A Note on the History of Mark-Recapture Population Estimates », *The Journal of Animal Ecology*, 34, p. 453 – 454.

Fellegi, I. P., et A. B. Sunter (1969), « A Theory of Record Linkage », *Journal of the American Statistical Association*, 64, p. 1183 – 1210.

Liu, B., L. Stokes, T. Topping, et G. Stunz (2017), « Estimation of a Total from a Population of Unknown Size and Application to Estimating Recreational Red Snapper Catch in Texas », *Journal of Survey Statistics and Methodology*, 100, p. 222 – 230.

National Research Council (2017), *Review of the Marine Recreational Information Program*, Washington: National Academies Press.

Tarnecki, J. H., et W. F. Patterson (2015), « Changes in Red Snapper Diet and Tropical Ecology Following the Deepwater Horizon Oil Spill », Marine and Coastal Fisheries, 7, p. 135 – 147.