aan SQS Team Rechtsbescherming en Veiligheid

cc

van Sander Scholtus

onderwerp classificatiefouten vuurwapenanalyse

datum 21 juni 2022

Inleiding

Team Rechtsbescherming en Veiligheid beschikt over een dataset U met circa 360 000 geregistreerde incidenten. Het doel is om binnen deze populatie het aantal vuurwapenincidenten te schatten. Definieer het label $s_i = 1$ als $i \in U$ in werkelijkheid een vuurwapenincident is en anders $s_i = 0$. De te schatten parameter is dan het aantal geregistreerde incidenten met label 1:

$$\theta = \sum_{i \in U} s_i. \tag{1}$$

Daarnaast is men geïnteresseerd in het aantal vuurwapenincidenten binnen subpopulaties van U op basis van achtergrondkenmerken van het incident, zoals leeftijd en geslacht van de betrokken verdachte.

In de praktijk zijn de werkelijke labels s_i voor bijna alle incidenten onbekend. In plaats daarvan zijn voorspelde labels beschikbaar, deels afkomstig van een *educated guess* en deels afkomstig uit een machine learning-algoritme (meer details volgen verderop). In de voorspelde labels komen classificatiefouten voor: sommige incidenten zijn ten onrechte geclassificeerd als vuurwapenincident en omgekeerd zijn sommige werkelijke vuurwapenincidenten niet als zodanig geclassificeerd. Uit steekproefsgewijze controles is een indicatie bekend van de verwachte percentages misclassificaties in beide richtingen, zowel voor de *educated guess* als voor de voorspellingen uit het machine learning-algoritme. Het doel van deze memo is om meer inzicht te geven in het effect van de classificatiefouten op de kwaliteit van het geschatte aantal vuurwapenincidenten, uitgedrukt in de vertekening en variantie van dit geschatte aantal. Verder gaan we in op de vraag hoe deze vertekening en variantie doorwerken in uitsplitsingen van het geschatte aantal naar subpopulaties.

Meer precies valt de populatie U op basis van de beschikbare informatie uiteen in twee strata: $U = A \cup B$. Voor de incidenten uit stratum A, met $|A| = 94\,800$, is een *educated guess* beschikbaar voor het echte label s_i, die we noteren als $\hat{s}_i^G$. Er geldt dat $\hat{s}_i^G = 1$ voor 30 000 incidenten in stratum A en $\hat{s}_i^G = 0$ voor de overige 64 800 incidenten. Voor de incidenten in stratum B is $\hat{s}_i^G$ niet beschikbaar. Voor deze incidenten is wel een voorspelling van $\hat{s}_i^G$ beschikbaar uit een machine learning-algoritme, genoteerd als $\hat{s}_i^M$. Dit algoritme is getraind op de waargenomen labels $\hat{s}_i^G$ in stratum A en

vervolgens toegepast op stratum B. Binnen stratum B geldt dat $\hat{s}_i^M = 1$ voor 54 300 incidenten en $\hat{s}_i^M = 0$ voor 211 400 incidenten.[1]

Per definitie is $\theta = \theta_A + \theta_B$, met $\theta_A = \sum_{i \in A} s_i$ en $\theta_B = \sum_{i \in B} s_i$. Gegeven de beschikbare informatie is een voor de hand liggende schatter voor θ gegeven door:

$$\hat{\theta}^{GM} = \hat{\theta}_A^G + \hat{\theta}_B^M = \sum_{i \in A} \hat{s}_i^G + \sum_{i \in B} \hat{s}_i^M. \qquad (2)$$

Voor de bovengenoemde aantallen is deze schatting gelijk aan $\hat{\theta}^{GM} = 30\,000 + 54\,300 = 84\,300$. Uitsplitsingen van $\hat{\theta}^{GM}$ naar subpopulaties van U op basis van achtergrondkenmerken kunnen worden gevonden door formule (2) te berekenen per subpopulatie.

Zoals eerder is opgemerkt komen in de waargenomen labels $\hat{s}_i^G$ en $\hat{s}_i^M$ toevallige classificatiefouten voor ten opzichte van de werkelijke labels s_i. Deze classificatiefouten veroorzaken variantie en mogelijk vertekening in de schatter $\hat{\theta}^{GM}$. In het vervolg van deze memo wordt uitgewerkt hoe groot deze vertekening en variantie waarschijnlijk zijn, uitgaande van de beschikbare steekproefinformatie over de percentages misclassificaties in $\hat{s}_i^G$ en $\hat{s}_i^M$.

Vertekening en variantie: theorie

Om de vertekening en variantie te bepalen van $\hat{\theta}^{GM}$ uit formule (2) als schatter voor θ uit formule (1), kunnen we de twee termen $\hat{\theta}_A^G$ en $\hat{\theta}_B^M$ apart beschouwen. Er geldt namelijk:

$$\begin{aligned} B(\hat{\theta}^{GM}) &= E(\hat{\theta}^{GM}) - \theta = E(\hat{\theta}_A^G + \hat{\theta}_B^M) - (\theta_A + \theta_B) = B(\hat{\theta}_A^G) + B(\hat{\theta}_B^M); \\ V(\hat{\theta}^{GM}) &= V(\hat{\theta}_A^G) + V(\hat{\theta}_B^M); \end{aligned} \qquad (3)$$

hier en in het vervolg geeft $B(.)$ de vertekening van een schatter aan, $E(.)$ de verwachte waarde en $V(.)$ de variantie. Met betrekking tot de variantie nemen we aan dat de incidenten onafhankelijk van elkaar zijn geclassificeerd.[2]

---

[1] Twee opmerkingen vooraf over de aantallen die in deze memo worden genoemd:

- In de informatie die voor deze memo beschikbaar was zijn alle tellingen afgerond op honderdtallen. Door deze afronding zijn soms kleine inconsistenties ontstaan; zo bevat stratum B volgens de brondata in totaal 265 800 incidenten in plaats van 54 300 + 211 400 = 265 700. In deze memo zijn steeds de afgeronde aantallen op het meest gedetailleerde niveau als uitgangspunt genomen; in berekeningen gebruiken we daarom bijvoorbeeld $|B| = 265\,700$.
- Verder is in de brondata sprake van 3 600 extra incidenten die met zekerheid als vuurwarenincident zijn geclassificeerd. Deze vallen buiten de populatie U zoals hier gedefinieerd, onder de aanname dat bij deze incidenten geen classificatiefouten voorkomen. Indien nodig kunnen deze 3 600 achteraf worden opgeteld bij de geschatte aantallen vuurwapenincidenten zoals vermeld in deze memo, al dan niet uitgesplitst naar subpopulatie.

[2] Het machine learning-algoritme waaruit $\hat{\theta}_B^M$ afkomstig is, is getraind op waargenomen labels $\hat{s}_i^G$ uit stratum A. Dit zou een afhankelijkheid kunnen introduceren tussen de schattingen over stratum A en B, en daarmee een covariantie tussen $\hat{\theta}_A^G$ en $\hat{\theta}_B^M$. In het vervolg gaan we ervan uit dat deze covariantie verwaarloosbaar klein is, omdat het een tweede-orde effect betreft (de voorspelde labels $\hat{s}_i^M$ zijn onafhankelijk toegewezen per incident i). Verder blijkt uit de resultaten later in deze memo dat de variantie van $\hat{\theta}^{GM}$ veel kleiner is dan de vertekening, die sowieso niet gevoelig is voor deze covariantie.

Aanpak op basis van misclassificatiekansen

Voor $\hat{\theta}_A^G$ geldt dat de vertekening en variantie worden bepaald door de kwaliteit van $\hat{s}_i^G$ als voorspelling voor s_i binnen stratum A. Beschouw de matrix

$$\mathbf{P}_A^G = \begin{pmatrix} p_{A,11}^G & 1 - p_{A,11}^G \\ 1 - p_{A,00}^G & p_{A,00}^G \end{pmatrix},$$

waarbij $p_{A,11}^G = P(\hat{s}_i^G = 1 | s_i = 1, i \in A)$ en $p_{A,00}^G = P(\hat{s}_i^G = 0 | s_i = 0, i \in A)$. Met behulp van deze misclassificatiekansen geldt voor de vertekening $B(\hat{\theta}_A^G)$ en voor de variantie $V(\hat{\theta}_A^G)$:

$$\begin{aligned} B(\hat{\theta}_A^G) &= E(\hat{\theta}_A^G) - \theta_A = \theta_A(p_{A,11}^G - 1) + (|A| - \theta_A)(1 - p_{A,00}^G); \\ V(\hat{\theta}_A^G) &= \theta_A p_{A,11}^G(1 - p_{A,11}^G) + (|A| - \theta_A)p_{A,00}^G(1 - p_{A,00}^G). \end{aligned} \tag{4}$$

Voor een afleiding van deze formules, zie Buonaccorsi (2010), Meertens et al. (2019) of Scholtus & van Delden (2020).

Voor $\hat{\theta}_B^M$ worden de vertekening en variantie bepaald door de kwaliteit van $\hat{s}_i^M$ als voorspelling voor s_i binnen stratum B. Beschouw de matrix

$$\mathbf{P}_B^M = \begin{pmatrix} p_{B,11}^M & 1 - p_{B,11}^M \\ 1 - p_{B,00}^M & p_{B,00}^M \end{pmatrix},$$

waarbij $p_{B,11}^M = P(\hat{s}_i^M = 1 | s_i = 1, i \in B)$ en $p_{B,00}^M = P(\hat{s}_i^M = 0 | s_i = 0, i \in B)$. Analoog aan formule (4) geldt:

$$\begin{aligned} B(\hat{\theta}_B^M) &= E(\hat{\theta}_B^M) - \theta_B = \theta_B(p_{B,11}^M - 1) + (|B| - \theta_B)(1 - p_{B,00}^M); \\ V(\hat{\theta}_B^M) &= \theta_B p_{B,11}^M(1 - p_{B,11}^M) + (|B| - \theta_B)p_{B,00}^M(1 - p_{B,00}^M). \end{aligned} \tag{5}$$

Alternatieve, stapsgewijze aanpak voor stratum B

Een alternatieve manier om de vertekening en variantie van $\hat{\theta}_B^M$ af te leiden is via een stapsgewijze aanpak:

- Bereken eerst de verwachting en variantie van $\hat{\theta}_B^M = \sum_{i \in B} \hat{s}_i^M$ als schatter voor $\hat{\theta}_B^G = \sum_{i \in B} \hat{s}_i^G$, de hypothetische schatter die zou zijn gevonden door $\hat{s}_i^G$ ook te tellen binnen stratum B.
- Bereken vervolgens de verwachting en variantie van de hypothetische schatter $\hat{\theta}_B^G$ als schatter voor θ_B, op dezelfde manier als in formule (4) voor $\hat{\theta}_A^G$.
- Combineer deze resultaten om de totale vertekening en variantie van $\hat{\theta}_B^M$ als schatter voor θ_B te bepalen.

Deze aanpak leidt tot de volgende uitdrukkingen voor de vertekening en variantie van $\hat{\theta}_B^M$; zie de appendix voor een afleiding.

$$
\begin{aligned}
B_{alt}(\hat{\theta}_B^M) &= |B|\{(1-p_{B,00}^G)p_{B,11}^{M|G} + p_{B,00}^G(1-p_{B,00}^{M|G})\} \\
&\quad + \theta_B\{(p_{B,11}^G + p_{B,00}^G - 1)(p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1) - 1\}; \\
V_{alt}(\hat{\theta}_B^M) &= |B|\{(1-p_{B,00}^G)p_{B,11}^{M|G}(1-p_{B,11}^{M|G}) + p_{B,00}^G p_{B,00}^{M|G}(1-p_{B,00}^{M|G})\} \\
&\quad + \theta_B(p_{B,11}^G + p_{B,00}^G - 1)\{p_{B,11}^{M|G}(1-p_{B,11}^{M|G}) - p_{B,00}^{M|G}(1-p_{B,00}^{M|G})\} \\
&\quad + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)^2\{\theta_B p_{B,11}^G(1-p_{B,11}^G) + (|B|-\theta_B)p_{B,00}^G(1-p_{B,00}^G)\}.
\end{aligned}
\tag{6}
$$

Hierbij is $p_{B,11}^G = P(\hat{s}_i^G = 1 | s_i = 1, i \in B)$ en $p_{B,00}^G = P(\hat{s}_i^G = 0 | s_i = 0, i \in B)$, analoog aan de eerder gedefinieerde kansen $p_{A,11}^G$ en $p_{A,00}^G$. Verder is $p_{B,11}^{M|G} = P(\hat{s}_i^M = 1 | \hat{s}_i^G = 1, i \in B)$ en $p_{B,00}^{M|G} = P(\hat{s}_i^M = 0 | \hat{s}_i^G = 0, i \in B)$.

Puur theoretisch zijn formules (5) en (6) equivalent aan elkaar. In de praktijk zou een voordeel van formule (6) kunnen zijn dat bij het schatten van deze uitdrukkingen ook informatie over de *recall* van het machine learning-algoritme kan worden meegenomen, in de vorm van de kansen $p_{B,11}^{M|G}$ en $p_{B,00}^{M|G}$. Zoals we in de volgende paragraaf zullen zien is over de recall van het machine learning-algoritme informatie beschikbaar uit een relatief grote aselecte steekproef, waarmee deze vrij nauwkeurig kan worden bepaald. Daarentegen is voor het schatten van de kansen $p_{B,11}^M$ en $p_{B,00}^M$ uit formule (5) informatie beschikbaar uit een veel kleinere steekproef die door experts is bekeken.

Aanpak op basis van calibratiekansen

Een mogelijk nadeel van alle bovenstaande formules voor vertekening en variantie is dat de (onbekende) echte aantallen θ_A en θ_B erin voorkomen. In de praktijk kunnen deze worden geschat door $\hat{\theta}_A^G$ en $\hat{\theta}_B^M$ – met als risico dat er vertekening optreedt in de geschatte vertekening en variantie – of direct uit een kleine steekproef die door experts is bekeken, wat echter leidt tot meer onzekerheid over de werkelijke vertekening en variantie. Een geheel andere mogelijke aanpak is om direct een voor vertekening *gecorrigeerde* schatter te berekenen, via zogenaamde calibratiekansen. Voor stratum A zijn de calibratiekansen gegeven door $c_{A,11}^G = P(s_i = 1 | \hat{s}_i^G = 1, i \in A)$ en $c_{A,00}^G = P(s_i = 0 | \hat{s}_i^G = 0, i \in A)$. Een onvertekende schatter voor θ_A is gegeven door:

$$
\hat{\theta}_A^{cal} = \hat{\theta}_A^G \hat{c}_{A,11}^G + (|A| - \hat{\theta}_A^G)(1 - \hat{c}_{A,00}^G),
\tag{7}
$$

waarbij $\hat{c}_{A,11}^G$ en $\hat{c}_{A,00}^G$ zijn geschat uit een steekproef die door experts is bekeken (zie ook de volgende paragraaf). Kloos et al. (2021) laten zien dat deze schatter onvertekend is, mits de steekproef waaruit de calibratiekansen zijn geschat representatief is voor de doelpopulatie.

De variantie van $\hat{\theta}_A^{cal}$ kan bij benadering worden geschat door [zie formule (2.8) in Buonaccorsi (2010)]:

$$
\hat{V}(\hat{\theta}_A^{cal}) = (\hat{\theta}_A^G)^2 \hat{V}(\hat{c}_{A,11}^G) + (|A| - \hat{\theta}_A^G)^2 \hat{V}(\hat{c}_{A,00}^G) + (\hat{c}_{A,11}^G + \hat{c}_{A,00}^G - 1)^2 \hat{V}(\hat{\theta}_A^G).
\tag{8}
$$

In deze formule zijn $\hat{V}(\hat{c}_{A,11}^G) = \hat{c}_{A,11}^G(1 - \hat{c}_{A,11}^G)/n_{A,+1}^G$ en $\hat{V}(\hat{c}_{A,11}^G) = \hat{c}_{A,00}^G(1 - \hat{c}_{A,00}^G)/n_{A,+0}^G$ de geschatte steekproefvarianties van $\hat{c}_{A,11}^G$ en $\hat{c}_{A,00}^G$, waarbij $n_{A,+1}^G$ het aantal gevallen met $\hat{s}_i^G = 1$ in de

steekproef is en $n^G_{A,+0}$ het aantal gevallen met $\hat{s}^G_i = 0$.[3] Verder is $\hat{V}(\hat{\theta}^G_A)$ een schatting voor $V(\hat{\theta}^G_A)$ uit (4). In de praktijk zijn de eerste twee termen uit (8) bepalend voor de nauwkeurigheid, omdat $V(\hat{\theta}^G_A)$ van een kleinere orde is dan $V(\hat{c}^G_{A,11})$ en $V(\hat{c}^G_{A,00})$.

Op dezelfde manier vinden we voor stratum B als calibratieschatter voor θ_B:

$$\hat{\theta}^{cal}_B = \hat{\theta}^M_B \hat{c}^M_{B,11} + (|B| - \hat{\theta}^M_B)(1 - \hat{c}^M_{B,00}), \tag{9}$$

waarbij de kansen $\hat{c}^M_{B,11}$ en $\hat{c}^M_{B,00}$ weer zijn geschat uit een steekproef die door experts is bekeken, met als bijbehorende geschatte variantie:

$$\hat{V}(\hat{\theta}^{cal}_B) = (\hat{\theta}^M_B)^2 \hat{V}(\hat{c}^M_{B,11}) + (|B| - \hat{\theta}^G_B)^2 \hat{V}(\hat{c}^M_{B,00}) + (\hat{c}^M_{B,11} + \hat{c}^M_{B,00} - 1)^2 \hat{V}(\hat{\theta}^M_B), \tag{10}$$

met $\hat{V}(\hat{c}^M_{B,11}) = \hat{c}^M_{B,11}(1 - \hat{c}^M_{B,11})/n^M_{B,+1}$ en $\hat{V}(\hat{c}^M_{B,00}) = \hat{c}^M_{B,00}(1 - \hat{c}^M_{B,00})/n^M_{B,+0}$.

Een schatter voor het totaal $\theta = \theta_A + \theta_B$ is nu gegeven door $\hat{\theta}^{cal} = \hat{\theta}^{cal}_A + \hat{\theta}^{cal}_B$. Onder de aanname dat $\hat{\theta}^{cal}_A$ en $\hat{\theta}^{cal}_B$ inderdaad onvertekend zijn kan de vertekening in $\hat{\theta}^{GM} = \hat{\theta}^G_A + \hat{\theta}^M_B$ nu worden geschat door:

$$\hat{B}(\hat{\theta}^{GM}) = \hat{B}(\hat{\theta}^G_A) + \hat{B}(\hat{\theta}^M_B) = (\hat{\theta}^G_A - \hat{\theta}^{cal}_A) + (\hat{\theta}^M_B - \hat{\theta}^{cal}_B) = \hat{\theta}^{GM} - \hat{\theta}^{cal}. \tag{11}$$

Het is ook mogelijk om een voor vertekening gecorrigeerde schatter te berekenen op basis van de misclassificatiekansen in $\mathbf{P}^G_A$ en $\mathbf{P}^M_B$ in plaats van calibratiekansen; zie bijvoorbeeld Buonaccorsi (2010). Hiermee zou, op dezelfde manier als in formule (11), de vertekening in $\hat{\theta}^{GM}$ kunnen worden geschat. Kloos et al. (2021) laten echter zien dat de gecorrigeerde schatter gebaseerd op misclassificatiekansen in de praktijk een grotere variantie heeft dan de calibratieschatter en dat hij instabiel kan worden als de kans op misclassificatie groot is. We zullen deze schatter daarom hier niet gebruiken.

Beschikbare informatie over de kwaliteit van het classificeren

Om de formules voor vertekening en variantie als gevolg van classificatiefouten uit de vorige paragraaf toe te kunnen passen is informatie nodig om de onbekende parameters in deze formules te schatten, zoals de foutkansen $p^M_{B,11}$ en $p^M_{B,00}$ en de calibratiekansen $c^G_{A,11}$ en $c^G_{A,00}$. Voor de classificatie van incidenten is dergelijke informatie beschikbaar uit twee bronnen.

Expertsteekproef

Ten eerste is voor een kleine steekproef van 330 incidenten geprobeerd om het echte label s_i te laten bepalen door experts. Van deze incidenten waren er 62 ouder dan vijf jaar, zodat de benodigde data niet langer toegankelijk waren en het niet mogelijk was om het echte label te

[3] In de afleiding van formule (8) en de formules voor $\hat{V}(\hat{c}^G_{A,11})$ en $\hat{V}(\hat{c}^G_{A,00})$ is door Buonaccorsi (2010) aangenomen dat $\hat{c}^G_{A,11}$ en $\hat{c}^G_{A,00}$ afkomstig zijn uit een enkelvoudig aselecte steekproef. In de huidige toepassing zullen deze kansen worden geschat uit een *gestratificeerde* aselecte steekproef (zie de volgende paragraaf). Omdat de stratificatie is gebaseerd op $\hat{s}^G_i$ heeft deze geen invloed op de geschatte calibratiekansen $\hat{c}^G_{A,11}$ en $\hat{c}^G_{A,00}$ (wel op de geschatte misclassificatiekansen). In het bijzonder blijft de afleiding van formule (8) valide en kunnen de varianties $V(\hat{c}^G_{A,11})$ en $V(\hat{c}^G_{A,00})$ op dezelfde manier worden geschat.

achterhalen. Over de resterende 268 incidenten was ook niet in alle gevallen voldoende informatie beschikbaar voor de experts om een label te bepalen, zoals te zien is in de onderstaande tabellen.

Van de 268 incidenten waren er 68 afkomstig uit stratum A. Op basis van deze incidenten is de volgende kruistabel gemaakt van het werkelijke label s_i met de *educated guess* $\hat{s}_i^G$.

Tabel 1: Kruistabel van s_i en $\hat{s}_i^G$ binnen een steekproef uit stratum A.

$s_i \setminus \hat{s}_i^G$	$\hat{s}_i^G = 1$	$\hat{s}_i^G = 0$	totaal
$s_i = 1$	13	12	25
$s_i = 0$	10	24	34
s_i kan niet worden bepaald	9	0	9
totaal	32	36	68

De overige 200 incidenten, uit stratum B, zijn gebruikt om een kruistabel te maken van het werkelijke label s_i met het door het machine learning-algoritme voorspelde label $\hat{s}_i^M$.

Tabel 2: Kruistabel van s_i en $\hat{s}_i^M$ binnen een steekproef uit stratum B.

$s_i \setminus \hat{s}_i^M$	$\hat{s}_i^M = 1$	$\hat{s}_i^M = 0$	totaal
$s_i = 1$	56	19	75
$s_i = 0$	52	36	88
s_i kan niet worden bepaald	32	5	37
totaal	140	60	200

Deze aantallen zijn niet afkomstig uit een volledig aselecte steekproef: er was sprake van een gestratificeerde steekproef naar de kolommen van Tabel 1 en Tabel 2. Dat wil zeggen: er zijn 32 willekeurige incidenten geselecteerd uit de groep met $\hat{s}_i^G = 1$ binnen stratum A, 36 willekeurige incidenten uit de groep met $\hat{s}_i^G = 0$ binnen stratum A, etc. (Dit is, voor zover we konden achterhalen, een redelijke benadering voor de manier waarop de steekproef tot stand is gekomen.)

Om de steekproefuitkomsten representatief te maken voor de doelpopulatie, kunnen we de aantallen in Tabel 1 en Tabel 2 wegen met de (bekende) insluitkans per steekproefstratum, zodat ze optellen tot de bekende kolomtotalen uit de populatie. Het resultaat van deze weging staat hieronder in Tabel 3 en Tabel 4:

Tabel 3: Gewogen kruistabel van s_i en $\hat{s}_i^G$ op basis van een steekproef uit stratum A.

$s_i \setminus \hat{s}_i^G$	$\hat{s}_i^G = 1$	$\hat{s}_i^G = 0$	totaal
$s_i = 1$	12 187,5	21 600,0	33 787,5
$s_i = 0$	9 375,0	43 200,0	52 575,0
s_i kan niet worden bepaald	8 437,5	0,0	8 437,5
totaal	30 000,0	64 800,0	94 800,0

Tabel 4: Gewogen kruistabel van s_i en $\hat{s}_i^M$ op basis van een steekproef uit stratum B.

$s_i \setminus \hat{s}_i^M$	$\hat{s}_i^M = 1$	$\hat{s}_i^M = 0$	totaal
$s_i = 1$	21 720,0	66 943,3	88 663,3
$s_i = 0$	20 168,6	126 840,0	147 008,6
s_i kan niet worden bepaald	12 411,4	17 616,7	30 028,1
totaal	54 300,0	211 400,0	265 700,0

Op basis van de bovenste twee rijen van Tabel 3 vinden we de volgende schatting voor de misclassificatiekansen $p_{A,11}^G$ en $p_{A,00}^G$ uit de matrix $\mathbf{P}_A^G$ in stratum A: $\hat{p}_{A,11}^G = 12\,187{,}5/33\,787{,}5 = 0{,}36$ en $\hat{p}_{A,00}^G = 43\,200/52\,575 = 0{,}82$. Voor stratum B vinden we, geheel analoog, op basis van de bovenste twee rijen van Tabel 4: $\hat{p}_{B,11}^M = 21\,720/88\,663{,}3 = 0{,}24$ en $\hat{p}_{B,00}^M = 126\,840/147\,008{,}6 = 0{,}86$. Volgens deze geschatte kansen komen in beide strata relatief veel classificatiefouten voor, waarbij er met name veel werkelijke vuurwapenincidenten ten onrechte worden gemist.

Kwaliteit machine learning-algoritme

Ten tweede is een aselecte steekproef van 18 234 incidenten uit stratum A (dat wil zeggen circa 20% van dit stratum) gebruikt om de *recall* en *precision* te bepalen van het machine learning-algoritme. Recall is de kans dat een werkelijk positief label correct wordt geclassificeerd door het algoritme; precision is de kans dat een positief voorspeld label een correcte voorspelling betreft. In dit geval zeggen deze maten iets over de kwaliteit van $\hat{s}_i^M$ als voorspelling van de *educated guess* $\hat{s}_i^G$, niet als voorspelling van het echte label s_i. Tabel 5 toont de kruistabel van correct en niet-correct voorspelde labels door het machine learning-algoritme binnen deze aselecte steekproef.

Tabel 5: Kruistabel van $\hat{s}_i^G$ en $\hat{s}_i^M$ op basis van een aselecte steekproef uit stratum A.

$\hat{s}_i^G \setminus \hat{s}_i^M$	$\hat{s}_i^M = 1$	$\hat{s}_i^M = 0$	totaal
$\hat{s}_i^G = 1$	4 447	1 121	5 568
$\hat{s}_i^G = 0$	414	12 252	12 666
totaal	4 861	13 373	18 234

Uit deze aantallen volgt: voor het voorspellen van het label $\hat{s}_i^G = 1$ is de recall 0,80 en de precision 0,91; voor het voorspellen van het label $\hat{s}_i^G = 0$ is de recall 0,97 en de precision 0,92. Deze recall- en precision-waarden liggen vrij dicht bij 1, en in het bijzonder dichter bij 1 dan de eerder geschatte kansen $\hat{p}_{B,11}^M$ en $\hat{p}_{B,00}^M$. Dit houdt in dat het algoritme vrij goed in staat is om de *educated guess*-labels te voorspellen. We hebben echter gezien dat de *educated guess*-labels relatief veel classificatiefouten bevatten. Zoals blijkt uit Tabel 4 zijn de recall en precision van het machine learning-algoritme voor het voorspellen van de *werkelijke* labels daarom een stuk lager dan Tabel 5 misschien suggereert.

Vertekening en variantie: toepassing

Schatten van de bijdrage van stratum A via misclassificatiekansen [formule (4)]
Om formule (4) toe te passen moeten de kansen uit de matrix $\mathbf{P}_A^G$ worden geschat en moet een (voorlopige) schatting voor het werkelijke aantal θ_A worden ingevuld. Voor het schatten van de matrix $\mathbf{P}_A^G$ maken we gebruik van Tabel 3 en vinden we, zoals eerder is beschreven, $\hat{p}_{A,11}^G = 0{,}36$ en $\hat{p}_{A,00}^G = 0{,}82$. Het onbekende werkelijke aantal θ_A zouden we kunnen schatten door $\hat{\theta}_A^G = 30\,000$, maar dit zou tot verkeerde conclusies kunnen leiden als $\hat{\theta}_A^G$ veel vertekening bevat (wat we juist willen onderzoeken). Uit voorzorg stellen we daarom voor om θ_A te schatten uit Tabel 3 gebaseerd op de kleine steekproef. Onder de aanname dat de incidenten waarvoor s_i niet bepaald kon worden een willekeurige deelsteekproef vormen, vinden we als schatting:

$$\hat{\theta}_A^{stp} = 94\,800 \times \frac{33\,787{,}5}{33\,787{,}5 + 52\,575} = 37\,088.$$

Invullen van $\hat{p}_{A,11}^G$, $\hat{p}_{A,00}^G$ en $\hat{\theta}_A^{stp}$ in formule (4) geeft:[4]

$$\hat{B}\left(\hat{\theta}_A^G\right) = 37\,088 \times (-0{,}64) + 57\,712 \times 0{,}18 = -13\,419;$$
$$\hat{V}\left(\hat{\theta}_A^G\right) = 37\,088 \times 0{,}36 \times 0{,}64 + 57\,712 \times 0{,}82 \times 0{,}18 = 17\,008 = (130{,}4)^2.$$

Merk op: de onzekerheid in de geschatte parameters $\hat{p}_{A,11}^G$, $\hat{p}_{A,00}^G$ en $\hat{\theta}_A^{stp}$ werkt door in de schatting van de vertekening en variantie van $\hat{\theta}_A^G$, maar heeft geen invloed op de echte vertekening of variantie van $\hat{\theta}_A^G$, aangezien deze geschatte parameters niet nodig zijn om $\hat{\theta}_A^G$ zelf te berekenen.

Schatten van de bijdrage van stratum A via calibratiekansen [formule (7)]
We bekijken ook de alternatieve aanpak gebaseerd op calibratiekansen. Op basis van de bovenste twee rijen van Tabel 3 schatten we: $\hat{c}_{A,11}^G = 12\,187{,}5/21\,562{,}5 = 0{,}57$ en $\hat{c}_{A,00}^G = 43\,200/64\,800 = 0{,}67$. (Hierbij gebruiken we wederom de aanname dat de incidenten waarvoor s_i niet bepaald kon worden een willekeurige deelsteekproef zijn.) De schatter $\hat{\theta}_A^{cal}$ uit formule (7) is nu gegeven door:

$$\hat{\theta}_A^{cal} = 30\,000 \times 0{,}57 + 64\,800 \times 0{,}33 = 38\,557.$$

De bijbehorende geschatte vertekening in $\hat{\theta}_A^G$ is gelijk aan $\hat{B}\left(\hat{\theta}_A^G\right) = 30\,000 - 38\,557 = -8\,557$.

Bij de interpretatie van dit resultaat moet echter wel rekening worden gehouden met de grote onzekerheid in $\hat{\theta}_A^{cal}$. Volgens formule (8) wordt de variantie van $\hat{\theta}_A^{cal}$ geschat door:

$$\hat{V}\left(\hat{\theta}_A^{cal}\right) = (30\,000)^2 \times \frac{0{,}57 \times 0{,}43}{23} + (64\,800)^2 \times \frac{0{,}67 \times 0{,}33}{36} + (0{,}23)^2 \times 17\,008$$
$$= 35\,537\,089 = (5\,961{,}3)^2.$$

Hierbij is op basis van de eerste twee rijen van Tabel 1 ingevuld dat $n_{A,+1}^G = 23$ en $n_{A,+0}^G = 36$. Verder is voor $\hat{V}\left(\hat{\theta}_A^G\right)$ de zojuist gevonden schatting op basis van misclassificatiekansen ingevuld.

[4] Voor alle berekeningen in deze paragraaf geldt dat ze zijn uitgevoerd zonder tussentijds af te ronden. Bij het weergeven van de berekeningen in deze memo zijn getallen waar nodig wel afgerond. Dit leidt soms tot schijnbare kleine inconsistenties.

Hieruit volgt dat de standaardfout van $\hat{\theta}_A^{cal}$ (en daarmee van $\hat{B}(\hat{\theta}_A^G)$) zo groot is dat niet kan worden geconcludeerd dat de vertekening in $\hat{\theta}_A^G$ significant verschilt van 0 (gebruikmakend van een 95%-betrouwbaarheidsinterval rond $\hat{\theta}_A^{cal}$ onder de aanname dat deze schatter een normale verdeling volgt). Dat de standaardfout van $\hat{\theta}_A^{cal}$ veel groter is dan die van $\hat{\theta}_A^G$, komt doordat $\hat{\theta}_A^{cal}$ gebruikmaakt van geschatte parameters uit de kleine steekproef, terwijl $\hat{\theta}_A^G$ los van deze steekproef is bepaald.

Schatten van de bijdrage van stratum B via misclassificatiekansen [formule (5)]
Om formule (5) toe te passen moeten de kansen uit de matrix $\mathbf{P}_B^M$ worden geschat en moet een (voorlopige) schatting voor het werkelijke aantal θ_B worden ingevuld. We gebruiken hier dezelfde aanpak als bij stratum A, maar nu gebruikmakend van de informatie uit Tabel 4 in plaats van Tabel 3. Zoals eerder beschreven vinden we dat $\hat{p}_{B,11}^M = 0{,}24$ en $\hat{p}_{B,00}^M = 0{,}86$. Verder berekenen we uit de steekproef als schatting voor θ_B:

$$\hat{\theta}_B^{stp} = 265\,700 \times \frac{88\,663{,}3}{88\,663{,}3 + 147\,008{,}6} = 99\,960.$$

Hiermee vinden we:

$$\hat{B}(\hat{\theta}_B^M) = 99\,960 \times (-0{,}76) + 165\,740 \times 0{,}14 = -52\,735;$$
$$\hat{V}(\hat{\theta}_B^M) = 99\,960 \times 0{,}24 \times 0{,}76 + 165\,740 \times 0{,}86 \times 0{,}14 = 38\,108 = (195{,}2)^2.$$

Schatten van de bijdrage van stratum B via calibratiekansen [formule (9)]
Voor stratum B kan ook weer gebruik worden gemaakt van calibratiekansen. Op basis van Tabel 4 schatten we, op dezelfde manier als boven voor stratum A: $\hat{c}_{B,11}^M = 21\,720/41\,888{,}6 = 0{,}52$ en $\hat{c}_{B,00}^M = 126\,840/193\,783{,}3 = 0{,}65$. Uit formule (9) volgt nu als schatter $\hat{\theta}_B^{cal}$:

$$\hat{\theta}_B^{cal} = 54\,300 \times 0{,}52 + 211\,400 \times 0{,}35 = 101\,185.$$

De bijbehorende geschatte vertekening in $\hat{\theta}_B^M$ is $\hat{B}(\hat{\theta}_B^M) = 54\,300 - 101\,185 = -46\,885$. De variantie van $\hat{\theta}_B^{cal}$ wordt volgens formule (10) geschat door:

$$\hat{V}(\hat{\theta}_B^{cal}) = (54\,300)^2 \times \frac{0{,}52 \times 0{,}48}{108} + (211\,400)^2 \times \frac{0{,}65 \times 0{,}35}{55} + (0{,}17)^2 \times 38\,108$$
$$= 190\,546\,109 = (13\,803{,}8)^2.$$

De standaardfout van $\hat{\theta}_B^{cal}$ (en daarmee van $\hat{B}(\hat{\theta}_B^M)$) is groot, maar desondanks zouden we op basis van een 95%-betrouwbaarheidsinterval rond $\hat{\theta}_B^{cal}$ (onder de aanname dat deze schatter een normale verdeling volgt) concluderen dat $\hat{\theta}_B^M$ een onderschatting geeft van het echte aantal θ_B.

Schatten van de bijdrage van stratum B via stapsgewijze aanpak [formule (6)]
Ten slotte, om gebruik te kunnen maken van formule (6) moeten we enkele aannames maken, omdat er geen directe informatie beschikbaar is over de kansen $p_{B,11}^{M|G}$, $p_{B,00}^{M|G}$, $p_{B,11}^G$ en $p_{B,00}^G$ die voorkomen in deze formule. De volgende twee aannames zijn voldoende:

1. De kwaliteit van $\hat{s}_i^M$ als voorspelling voor $\hat{s}_i^G$ is hetzelfde binnen strata A en B. Specifiek nemen we aan dat de kansen $p_{B,11}^{M|G} = P(\hat{s}_i^M = 1 | \hat{s}_i^G = 1, i \in B)$ en $p_{B,00}^{M|G} =$

$P(\hat{s}_i^M = 0 | \hat{s}_i^G = 0, i \in B)$ gelijk zijn aan de corresponderende kansen $p_{A,11}^{M|G} = P(\hat{s}_i^M = 1 | \hat{s}_i^G = 1, i \in A)$ en $p_{A,00}^{M|G} = P(\hat{s}_i^M = 0 | \hat{s}_i^G = 0, i \in A)$.

2. De kwaliteit van $\hat{s}_i^G$ als voorspelling voor s_i is hetzelfde binnen strata A en B. Specifiek nemen we aan dat de kansen $p_{B,11}^G = P(\hat{s}_i^G = 1 | s_i = 1, i \in B)$ en $p_{B,00}^G = P(\hat{s}_i^G = 0 | s_i = 0, i \in B)$ gelijk zijn aan de corresponderende kansen $p_{A,11}^G = P(\hat{s}_i^G = 1 | s_i = 1, i \in A)$ en $p_{A,00}^G = P(\hat{s}_i^G = 0 | s_i = 0, i \in A)$.

Beide aannames zijn in de praktijk niet te verifiëren, aangezien $\hat{s}_i^G$ alleen is waargenomen binnen stratum A. De waarden $\hat{s}_i^G$ binnen stratum B waarover in deze aannames wordt gesproken zijn puur hypothetisch.

Uit de aselecte steekproef van 18 234 incidenten uit stratum A is geschat dat $\hat{p}_{A,11}^{M|G} = 0{,}80$ (recall voor label 1) en $\hat{p}_{A,00}^{M|G} = 0{,}97$ (recall voor label 0). De kansen $p_{A,11}^G$ en $p_{A,00}^G$ worden, als voorheen, geschat door $\hat{p}_{A,11}^G = 0{,}36$ en $\hat{p}_{A,00}^G = 0{,}82$. De enige andere onbekende grootheid in formule (6) is θ_B; hiervoor wordt opnieuw $\hat{\theta}_B^{stp} = 99\,960$ ingevuld.

Gebruikmakend van de aannames dat $p_{B,11}^{M|G} = p_{A,11}^{M|G}$, $p_{B,00}^{M|G} = p_{A,00}^{M|G}$, $p_{B,11}^G = p_{A,11}^G$ en $p_{B,00}^G = p_{A,00}^G$ vinden we nu de volgende uitkomsten:

$$\begin{aligned}\hat{B}_{alt}(\hat{\theta}_B^M) &= 265\,700 \times (0{,}18 \times 0{,}80 + 0{,}82 \times 0{,}03) + 99\,960 \times (0{,}18 \times 0{,}77 - 1) \\ &= -41\,019; \\ \hat{V}_{alt}(\hat{\theta}_B^M) &= 265\,700 \times (0{,}18 \times 0{,}80 \times 0{,}20 + 0{,}82 \times 0{,}97 \times 0{,}03) \\ &\quad + \{99\,960 \times 0{,}18 \times (0{,}80 \times 0{,}20 - 0{,}97 \times 0{,}03)\} \\ &\quad + \{(0{,}77)^2 \times (99\,960 \times 0{,}36 \times 0{,}64 + 165\,740 \times 0{,}82 \times 0{,}18)\} \\ &= 44\,649 = (211{,}3)^2.\end{aligned}$$

Schatten van de totale vertekening en variantie

Met behulp van formule (3) en (11) volgen uit de bovenstaande resultaten ook schattingen voor de vertekening en variantie van de totale schatter $\hat{\theta}^{GM} = \hat{\theta}_A^G + \hat{\theta}_B^M$. Tabel 6 geeft een overzicht van de resultaten die worden gevonden bij de verschillende aanpakken die hier zijn gebruikt.

Tabel 6: Geschatte vertekening en variantie van het aantal vuurwapenincidenten volgens drie verschillende aanpakken.

aanpak		stratum A	stratum B	totaal
misclassificatiekansen in beide strata	vertekening	–13 419	–52 735	–66 154
	std.fout	130	195	235
misclassificatiekansen in stratum A, stapsgewijze aanpak in stratum B	vertekening	–13 419	–41 019	–54 438
	std.fout	130	211	248
calibratiekansen in beide strata	vertekening	–8 557	–46 885	–55 441
	std.fout	-	-	-

Wat de vertekening betreft komen de resultaten op basis van alle drie de aanpakken redelijk goed overeen. Op basis van deze uitkomsten zou de conclusie zijn dat het geschatte aantal $\hat{\theta}^{GM} = 84\,300$ waarschijnlijk een sterke onderschatting geeft van het werkelijke aantal θ. Het werkelijke aantal zou

eerder in de buurt liggen van 140 000 à 150 000. Er treedt onderschatting op in beide strata, maar het grootste deel komt voor rekening van stratum B, waar $\hat{s}_i^G$ onbekend is en de schatting gebaseerd wordt op de voorspellingen $\hat{s}_i^M$ uit het machine learning-algoritme.

Het verschil tussen de uitkomsten voor de verschillende aanpakken valt binnen de 95%-betrouwbaarheidsmarge (gebaseerd op een normale verdeling) van de calibratieschatter voor het werkelijke aantal waaruit de onderste resultaten in Tabel 6 zijn afgeleid. Deze calibratieschatter is gelijk aan $\hat{\theta}_A^{cal} + \hat{\theta}_B^{cal} = 38\,557 + 101\,185 = 139\,741$, met een bijbehorende standaardfout van $\sqrt{(5\,961{,}3)^2 + (13\,803{,}8)^2} = 15\,036{,}1$.

Vergeleken met de geschatte vertekening lijkt de standaardfout van $\hat{\theta}^{GM}$ als gevolg van misclassificaties verwaarloosbaar klein. De verschillende aanpakken zijn het redelijk met elkaar eens over de orde van grootte van deze standaardfout. (De aanpak op basis van calibratiekansen levert geen eigen schatting op voor de standaardfout van $\hat{\theta}^{GM}$, alleen een schatting voor de standaardfout van $\hat{\theta}_A^{cal} + \hat{\theta}_B^{cal}$.)

Subpopulaties

Stel nu dat de populatie U is opgedeeld in een aantal subpopulaties op basis van een achtergrondkenmerk X. Noteer U_x voor de subpopulatie van incidenten waarvoor $X = x$. We nemen aan dat $U = \bigcup_{x=1}^{K} U_x$, waarbij K het aantal verschillende klassen van X is. Op dezelfde manier noteren we $A = \bigcup_{x=1}^{K} A_x$ en $B = \bigcup_{x=1}^{K} B_x$.

Er geldt:

$$\theta = \sum_{i \in U} s_i = \sum_{x=1}^{K} \sum_{i \in U_x} s_i = \sum_{x=1}^{K} \theta_x ;$$

$$\hat{\theta}^{GM} = \hat{\theta}_A^G + \hat{\theta}_B^M = \sum_{i \in A} \hat{s}_i^G + \sum_{i \in B} \hat{s}_i^M = \sum_{x=1}^{K} \sum_{i \in A_x} \hat{s}_i^G + \sum_{x=1}^{K} \sum_{i \in B_x} \hat{s}_i^M = \sum_{x=1}^{K} \left(\hat{\theta}_{Ax}^G + \hat{\theta}_{Bx}^M\right).$$

Noteer verder: $\theta_{Ax} = \sum_{i \in A_x} s_i$ en $\theta_{Bx} = \sum_{i \in B_x} s_i$.

Voor de vertekening en variantie van $\hat{\theta}_{Ax}^G$, $\hat{\theta}_{Bx}^M$ en $\hat{\theta}_x^{GM} = \hat{\theta}_{Ax}^G + \hat{\theta}_{Bx}^M$ in subpopulatie U_x kunnen formules analoog aan (3) tot en met (11) worden afgeleid, met als verschil dat de foutkansen nu van toepassing zijn op alleen de betreffende subpopulatie. Gemakshalve werken we dit hier alleen uit voor formules (3), (4) en (5). Er geldt dat $B\left(\hat{\theta}_x^{GM}\right) = B\left(\hat{\theta}_{Ax}^G\right) + B\left(\hat{\theta}_{Bx}^M\right)$ en $V\left(\hat{\theta}_x^{GM}\right) = V\left(\hat{\theta}_{Ax}^G\right) + V\left(\hat{\theta}_{Bx}^M\right)$, waarbij:

$$\begin{aligned}
B\left(\hat{\theta}_{Ax}^G\right) &= \theta_{Ax}\left(p_{Ax,11}^G - 1\right) + \left(|A_x| - \theta_{Ax}\right)\left(1 - p_{Ax,00}^G\right); \\
V\left(\hat{\theta}_{Ax}^G\right) &= \theta_{Ax} p_{Ax,11}^G\left(1 - p_{Ax,11}^G\right) + \left(|A_x| - \theta_{Ax}\right) p_{Ax,00}^G\left(1 - p_{Ax,00}^G\right); \\
B\left(\hat{\theta}_{Bx}^M\right) &= \theta_{Bx}\left(p_{Bx,11}^M - 1\right) + \left(|B_x| - \theta_{Bx}\right)\left(1 - p_{Bx,00}^M\right); \\
V\left(\hat{\theta}_{Bx}^M\right) &= \theta_{Bx} p_{Bx,11}^M\left(1 - p_{Bx,11}^M\right) + \left(|B_x| - \theta_{Bx}\right) p_{Bx,00}^M\left(1 - p_{Bx,00}^M\right).
\end{aligned} \tag{12}$$

Er is nu een aantal verschillende situaties denkbaar:

1. Het achtergrondkenmerk X hangt niet samen met het werkelijke label s_i en ook niet met de waargenomen labels $\hat{s}_i^G$ en $\hat{s}_i^M$. In dit geval geldt dat $\theta_{Ax} = \frac{|A_x|}{|A|}\theta_A$ en $\theta_{Bx} = \frac{|B_x|}{|B|}\theta_B$ voor elke klasse x, aangezien de verdeling van s_i niet afhangt van het achtergrondkenmerk. Ook geldt dat $p_{Ax,11}^G = p_{A,11}^G$, $p_{Ax,00}^G = p_{A,00}^G$, $p_{Bx,11}^M = p_{B,11}^M$ en $p_{Bx,00}^M = p_{B,00}^M$, aangezien de classificatiefouten evenmin samenhangen met het achtergrondkenmerk. Uit formule (12) volgt nu dat:

$$B\left(\hat{\theta}_{Ax}^G\right) = \frac{|A_x|}{|A|}B\left(\hat{\theta}_A^G\right), \quad V\left(\hat{\theta}_{Ax}^G\right) = \frac{|A_x|}{|A|}V\left(\hat{\theta}_A^G\right),$$
$$B\left(\hat{\theta}_{Bx}^M\right) = \frac{|B_x|}{|B|}B\left(\hat{\theta}_B^M\right), \quad V\left(\hat{\theta}_{Bx}^M\right) = \frac{|B_x|}{|B|}V\left(\hat{\theta}_B^M\right).$$

   Dat wil zeggen: de vertekening en variantie per subpopulatie zijn evenredig aan de vertekening en variantie voor de totale populatie, waarbij de verhouding wordt bepaald door de relatieve omvang van de subpopulatie. Gegeven geschatte waarden voor $B\left(\hat{\theta}_A^G\right)$, $B\left(\hat{\theta}_B^M\right)$, $V\left(\hat{\theta}_A^G\right)$ en $V\left(\hat{\theta}_B^M\right)$ kan de corresponderende geschatte vertekening en variantie per subpopulatie direct worden afgeleid uit de relatieve omvang per subpopulatie.
   N.B. Voor de strata A en B afzonderlijk mogen we in dit geval altijd concluderen dat de vertekening in alle subpopulaties hetzelfde teken heeft. Voor de populatie U als geheel hoeft dit in het algemeen niet zo te zijn, aangezien de verhoudingen $|A_x|/|A|$ en $|B_x|/|B|$ niet per se gelijk aan elkaar zijn en $B\left(\hat{\theta}_A^G\right)$ en $B\left(\hat{\theta}_B^M\right)$ een tegengesteld teken kunnen hebben. In de huidige toepassing op vuurwapenincidenten lijken $B\left(\hat{\theta}_A^G\right)$ en $B\left(\hat{\theta}_B^M\right)$ echter wel hetzelfde teken te hebben (zie Tabel 6), en in dat geval is de conclusie dat de vertekening in alle subpopulaties hetzelfde teken heeft wel gerechtvaardigd.
2. Het achtergrondkenmerk X hangt mogelijk samen met het werkelijke label s_i maar, gegeven dit label, niet met de classificatiefouten in de waargenomen labels $\hat{s}_i^G$ en $\hat{s}_i^M$. In dit geval geldt nog steeds dat $p_{Ax,11}^G = p_{A,11}^G$, $p_{Ax,00}^G = p_{A,00}^G$, $p_{Bx,11}^M = p_{B,11}^M$ en $p_{Bx,00}^M = p_{B,00}^M$. Echter, de aantallen θ_{Ax} en θ_{Bx} zijn nu niet noodzakelijk evenredig aan θ_A en θ_B. De vertekening en variantie per subpopulatie zijn nu niet direct te relateren aan de vertekening en variantie voor de totale populatie. Om $B\left(\hat{\theta}_x^{GM}\right)$ en $V\left(\hat{\theta}_x^{GM}\right)$ te bepalen kan formule (12) worden gebruikt, waarbij het niet nodig is om de foutkansen apart te schatten voor elke subpopulatie. Wel moeten voor elke subpopulatie aparte schattingen van θ_{Ax} en θ_{Bx} worden ingevuld.
   Zolang het aantal subpopulaties niet al te groot is en de verdeling over deze subpopulaties niet al te scheef is, zouden schattingen voor θ_{Ax} en θ_{Bx} nog steeds kunnen worden afgeleid uit de steekproef van incidenten die door experts zijn bekeken; dat wil zeggen: uit de rijtotalen van Tabel 3 en Tabel 4, maar nu berekend per subpopulatie. Als het aantal beschikbare waarnemingen in deze steekproef voor bepaalde subpopulaties erg klein wordt, is het waarschijnlijk beter om de (mogelijk vertekende) populatieschattingen $\hat{\theta}_{Ax}^G$ en $\hat{\theta}_{Bx}^M$ te gebruiken. Een veiligere optie, indien haalbaar, is om de steekproef uit te breiden.
3. Het achtergrondkenmerk X hangt samen met de classificatiefouten in de waargenomen labels $\hat{s}_i^G$ en $\hat{s}_i^M$ en mogelijk ook met het werkelijke label s_i. In dit geval kan formule (12) niet verder worden vereenvoudigd. Om $B\left(\hat{\theta}_x^{GM}\right)$ en $V\left(\hat{\theta}_x^{GM}\right)$ te bepalen kan deze formule worden gebruikt, maar de foutkansen $p_{Ax,11}^G$, $p_{Ax,00}^G$, $p_{Bx,11}^M$ en $p_{Bx,00}^M$ moeten nu apart worden geschat voor elke subpopulatie. Ook moeten voor elke subpopulatie aparte

schattingen van θ_{Ax} en θ_{Bx} worden ingevuld.
In dit geval loopt men bij het schatten van de foutkansen $p^G_{Ax,11}$, $p^G_{Ax,00}$, $p^M_{Bx,11}$ en $p^M_{Bx,00}$ al snel aan tegen de beperkingen van de kleine steekproef die door experts is bekeken. Om voor elke subpopulatie een goede schatting van deze kansen te krijgen zou deze steekproef waarschijnlijk moeten worden uitgebreid. (Merk op: anders dan de schattingen voor de aantallen θ_{Ax} en θ_{Bx} die bij situatie 2 zijn genoemd maken de geschatte foutkansen ook gebruik van het binnenwerk van Tabel 3 en Tabel 4.)
Voor de bijdrage van stratum B is de alternatieve aanpak gebaseerd op formule (6) in plaats van (5) hier mogelijk aantrekkelijker, omdat daarbij minder foutkansen uit de kleine steekproef hoeven te worden afgeleid; in plaats daarvan moet de recall van het machine learning-algoritme per subpopulatie worden geschat, waarvoor een veel grotere steekproef beschikbaar is. Wel wordt bij deze alternatieve aanpak geleund op de eerder genoemde onverifieerbare aannames over de hypothetische labels $\hat{s}^G_i$ in stratum B (en nu per subpopulatie). Ook biedt deze aanpak geen oplossing voor het bepalen van de foutkansen $p^G_{Ax,11}$ en $p^G_{Ax,00}$.

Conclusie

In deze memo is op basis van een aantal verschillende aanpakken geprobeerd om een schatting te maken van de vertekening en variantie van een eerder geschat aantal vuurwapenincidenten. Het geschatte aantal $\hat{\theta}^{GM} = 84\,300$ is gebaseerd op een classificatie van geregistreerde incidenten, deels volgens een *educated guess* en deels volgens een machine learning-algoritme. Uit de resultaten die hier zijn afgeleid volgt dat dit cijfer hoogstwaarschijnlijk een sterke onderschatting geeft van het werkelijke aantal vuurwapenincidenten. Volgens de aanpakken die hier zijn bekeken zou het werkelijke aantal eerder in de buurt van $140\,000$ à $150\,000$ liggen. In vergelijking met de geschatte vertekening lijkt de standaardfout van het geschatte aantal $\hat{\theta}^{GM}$ verwaarloosbaar klein (235 volgens de eerste aanpak uit Tabel 6).

Met behulp van geschatte calibratiekansen is een voor vertekening gecorrigeerde schatter afgeleid voor het aantal vuurwapenincidenten. Deze schatter is gelijk aan $\hat{\theta}^{cal}_A + \hat{\theta}^{cal}_B = 139\,741$, met een standaardfout van $15\,036$ (een factor 60 groter dan de standaardfout vóór correctie). Deze laatste schatter is zuiver, mits de gewogen steekproef waaruit de calibratiekansen zijn geschat mag worden beschouwd als representatief voor de doelpopulatie. Als niet aan deze aanname voldaan is, is deze 'gecorrigeerde' schatter zelf mogelijk sterk vertekend; cf. Meertens et al. (2022). In dit geval lijkt de aanname van een representatieve steekproef (na weging) wel gerechtvaardigd. Een aandachtspunt is dat de oorspronkelijk getrokken steekproef een aantal incidenten bevat waarvoor de werkelijke classificatie niet kon worden bepaald, waarvan hier is aangenomen dat deze een willekeurige deelsteekproef vormen.

De grote onderschatting van het werkelijke aantal vuurwapenincidenten kan worden verklaard omdat in de gebruikte data relatief veel misclassificaties lijken voor te komen. Dit geldt zowel voor de *educated guess* als voor het machine learning-algoritme. Het algoritme veroorzaakt het grootste

deel van de onderschatting, al komt dit mede doordat het is toegepast op een relatief groot deel van de populatie. Voor de *educated guess* is uit de steekproef geschat dat slechts 36% van de werkelijke vuurwapenincidenten als zodanig is geclassificeerd, terwijl omgekeerd 82% van de werkelijke niet-vuurwapenincidenten correct is geclassificeerd. Voor het machine learning-algoritme zijn deze percentages respectievelijk 24% en 86%.

Appendix: afleiding van formule (6)

Noteer $E(\hat{\theta}_B^M|\hat{\theta}_B^G)$ en $V(\hat{\theta}_B^M|\hat{\theta}_B^G)$ voor de conditionele verwachting en variantie van $\hat{\theta}_B^M$ gegeven de hypothetische schatter $\hat{\theta}_B^G$. Voor de vertekening van $\hat{\theta}_B^M$ als schatter voor θ_B geldt:

$$\begin{aligned}
B(\hat{\theta}_B^M) &= E(\hat{\theta}_B^M) - \theta_B \\
&= E\{E(\hat{\theta}_B^M|\hat{\theta}_B^G)\} - \theta_B \\
&= E\{\hat{\theta}_B^G p_{B,11}^{M|G} + (|B| - \hat{\theta}_B^G)(1 - p_{B,00}^{M|G})\} - \theta_B \\
&= |B|(1 - p_{B,00}^{M|G}) + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)E(\hat{\theta}_B^G) - \theta_B \\
&= |B|(1 - p_{B,00}^{M|G}) + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)\{\theta_B p_{B,11}^G + (|B| - \theta_B)(1 - p_{B,00}^G)\} - \theta_B \\
&= |B|\{(1 - p_{B,00}^G)p_{B,11}^{M|G} + p_{B,00}^G(1 - p_{B,00}^{M|G})\} \\
&\qquad + \theta_B\{(p_{B,11}^G + p_{B,00}^G - 1)(p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1) - 1\}.
\end{aligned}$$

In de derde en vijfde regel zijn $E(\hat{\theta}_B^M|\hat{\theta}_B^G)$ en $E(\hat{\theta}_B^G)$ uitgewerkt op dezelfde manier die ook is gebruikt in formules (4) en (5).

Om de variantie van $\hat{\theta}_B^M$ te evalueren gebruiken we de klassieke aanpak om een totale variantie op te delen in conditionele componenten (Knottnerus, 2003, p. 14):

$$\begin{aligned}
V(\hat{\theta}_B^M) &= E\{V(\hat{\theta}_B^M|\hat{\theta}_B^G)\} + V\{E(\hat{\theta}_B^M|\hat{\theta}_B^G)\} \\
&= E\{\hat{\theta}_B^G p_{B,11}^{M|G}(1 - p_{B,11}^{M|G}) + (|B| - \hat{\theta}_B^G)p_{B,00}^{M|G}(1 - p_{B,00}^{M|G})\} \\
&\qquad + V\{\hat{\theta}_B^G p_{B,11}^{M|G} + (|B| - \hat{\theta}_B^G)(1 - p_{B,00}^{M|G})\} \\
&= |B|p_{B,00}^{M|G}(1 - p_{B,00}^{M|G}) + \{p_{B,11}^{M|G}(1 - p_{B,11}^{M|G}) - p_{B,00}^{M|G}(1 - p_{B,00}^{M|G})\}E(\hat{\theta}_B^G) \\
&\qquad + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)^2 V(\hat{\theta}_B^G) \\
&= |B|p_{B,00}^{M|G}(1 - p_{B,00}^{M|G}) \\
&\qquad + \{p_{B,11}^{M|G}(1 - p_{B,11}^{M|G}) - p_{B,00}^{M|G}(1 - p_{B,00}^{M|G})\}\{\theta_B p_{B,11}^G + (|B| - \theta_B)(1 - p_{B,00}^G)\} \\
&\qquad + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)^2\{\theta_B p_{B,11}^G(1 - p_{B,11}^G) + (|B| - \theta_B)p_{B,00}^G(1 - p_{B,00}^G)\} \\
&= |B|\{(1 - p_{B,00}^G)p_{B,11}^{M|G}(1 - p_{B,11}^{M|G}) + p_{B,00}^G p_{B,00}^{M|G}(1 - p_{B,00}^{M|G})\} \\
&\qquad + \theta_B(p_{B,11}^G + p_{B,00}^G - 1)\{p_{B,11}^{M|G}(1 - p_{B,11}^{M|G}) - p_{B,00}^{M|G}(1 - p_{B,00}^{M|G})\} \\
&\qquad + (p_{B,11}^{M|G} + p_{B,00}^{M|G} - 1)^2\{\theta_B p_{B,11}^G(1 - p_{B,11}^G) + (|B| - \theta_B)p_{B,00}^G(1 - p_{B,00}^G)\}.
\end{aligned}$$

Voor het evalueren van $E(\hat{\theta}_B^M|\hat{\theta}_B^G)$, $V(\hat{\theta}_B^M|\hat{\theta}_B^G)$, $E(\hat{\theta}_B^G)$ en $V(\hat{\theta}_B^G)$ in deze afleiding zijn weer dezelfde eigenschappen gebruikt als in formules (4) en (5).

Referenties

J.P. Buonaccorsi (2010), *Measurement Error: Models, Methods, and Applications*. Chapman & Hall/CRC, Boca Raton.

K. Kloos, Q.A. Meertens, S. Scholtus & J. Karch (2021), Comparing Correction Methods to Reduce Misclassification Bias. In: L. Cao, W. Kosters & J. Lijffijt (red.), *Proceedings of BNAIC/BENELEARN*, Springer, Leiden, pp. 103–129 (https://bnaic.liacs.leidenuniv.nl/wordpress/wp-content/uploads/papers/BNAICBENELEARN_2020_Final_paper_64.pdf).

P. Knottnerus (2003), *Sample survey theory: some Pythagorean perspectives*. Springer, New York.

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik & F.W. Takes (2019), A Bayesian approach for accurate classification-based aggregates. In: T.Y. Berger-Wolf & N.V. Chawla (red.), *Proceedings of the 19th SIAM International Conference on Data Mining (SDM)*, Calgary, pp. 306–314.

Q.A. Meertens, C.G.H. Diks, H.J. van den Herik & F.W. Takes (2022), Improving the output quality of official statistics based on machine learning algorithms. *Journal of Official Statistics* 38(2), pp. 1–25.

S. Scholtus & A. van Delden (2020), *The accuracy of estimators based on a binary classifier*. Discussion paper 202007, CBS, Den Haag.