DOOR KATRIEN ANTONIO

ONDER PROFESSOREN

Katrien Antonio

Katrien Antonio PhD is gewoon hoogleraar in actuariële wetenschappen aan KU Leuven en hoofddocent aan de Universiteit van Amsterdam. Zij is 'Actuaris van het jaar 2021'. Haar onderzoek concentreert zich rond data science methoden voor het actuariaat, met toepassingen in reservering, sterftemodellering, fraudedetectie en tarifering. Haar onderwijsportfolio bestaat uit bachelor, master, executive master en in company cursussen en dekt zowel de fundamenten van actuariële wetenschappen af alsook recente ontwikkelingen in het actuariële onderzoek.

Het modelleren van het aantal nameldingen is een bekend actuarieel probleem in het kader van schadereservering. Hierbij wordt de actuaris geconfronteerd met het schatten van het aantal schades dat zich voordeed in het verleden, maar gemeld zal worden in de toekomst. De zogeheten meldingsvertraging is de tijd die verloopt tussen het zich voordoen van de verzekerde gebeurtenis of het ongeval en het melden van de schade aan de verzekeraar. Een gepaste hoeveelheid kapitaal dient geboekt voor deze IBNR, oftewel Incurred But Not Reported, schades en het accuraat inschatten van het aantal van dergelijke schades is daarbij een eerste, belangrijke stap.

In ons recente artikel '*Modelling the occurrence of events subject to a reporting delay via an EM algorithm*'[1] laten we zien dat het modelleren van nameldingen een uitdaging is in heel wat wetenschappelijke disciplines. Samen met mijn coauteurs[2] structureer ik de multi-disciplinaire literatuur rond dit onderwerp, zodat overeenkomsten en verschillen tussen, en sterktes en zwaktes van voorgestelde modellen, duidelijk worden. Bovendien stellen we een nieuwe aanpak voor om nameldingen te modelleren, via het gezamenlijk schatten van een regressiemodel voor het zich voordoen van gebeurtenissen enerzijds en de meldingsvertraging anderzijds. Hierbij levert de inzet van covariaten bijkomende en verfijnde inzichten in het aantal nameldingen. Via de inzet van een Expectation-Maximization algoritme reduceren we het schattingsprobleem op een slimme manier tot het (herhaaldelijk) schatten van meer eenvoudige regressiemodellen. Op die manier creëert ons onderzoek een elegante, eenvoudig te begrijpen en te implementeren methode voor het voorspellen van het aantal nameldingen.

EEN HEDENDAAGSE TERM VOOR DIT STATISTISCHE PROBLEEM IS NOWCASTING

NOWCASTING

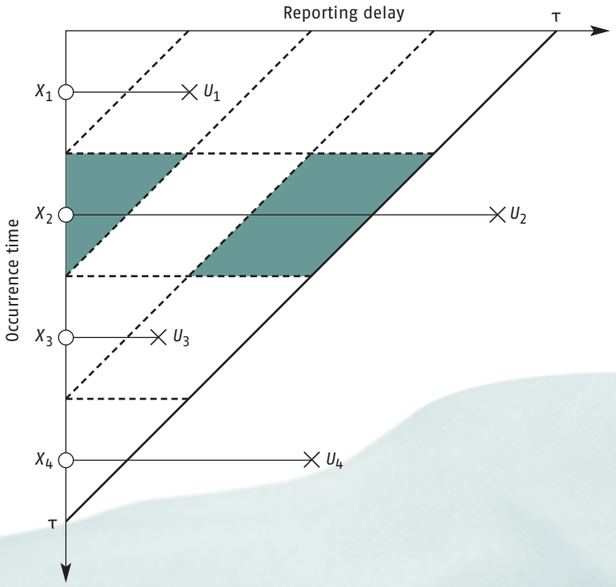
Ons artikel bekijkt het statistisch probleem waarbij een individu of object twee gebeurtenissen ervaart. Een eerste, of primaire, gebeurtenis doet zich voor op tijdstip x en een tweede, secundaire gebeurtenis doet zich voor op een later tijdstip s, met $s \geq x$. De vertraging $u = s-x$ tussen beide gebeurtenissen zorgt voor statistische uitdagingen, want het aantal gebeurtenissen dat geobserveerd wordt op een zeker tijdstip – zeg τ – is rechts-gecensureerd. Inderdaad, het aantal waargenomen gebeurtenissen is een onderschatting van het

aantal gebeurtenissen dat zich werkelijk tot dusver heeft gemanifesteerd. Anders bekeken is de meldingsvertraging onderhevig aan rechtse truncatie, want we observeren deze vertraging slechts als zij kleiner is dan de tijd tussen τ, het huidige evaluatiemoment, en het tijdstip x.

Dit statistische probleem gaat terug tot eind jaren '80 – begin jaren '90 met bijdragen van bijvoorbeeld Jewell in de actuariële literatuur en heel wat bijdragen in de biostatistische literatuur, in het bijzonder met betrekking tot het schatten van het aantal HIV-infecties. Een hedendaagse term voor dit statistische probleem is *nowcasting*, waarbij het aantal primaire gebeurtenissen geschat dient te worden aan de hand van de geregistreerde secundaire gebeurtenissen. Voorbeelden liggen voor het oprapen in diverse disciplines: het aantal IBNR-schades in het actuariaat, het aantal garantie ('warranty') claims in kwaliteitscontrole en diverse epidemiologische toepassingen zoals het monitoren van een uitbraak van een infectieziekte. In dat geval is x het tijdstip van diagnose en s het moment waarop de bevoegde instantie geïnformeerd wordt over de bevestigde case.

MODELLEN IN CONTINUE EN DISCRETE TIJD

Figuur 1 visualiseert vier gebeurtenissen en hun bijhorende meldingsvertraging. We zien dat observaties 1 en 3 volledig zijn, en gemeld worden voor het huidige evaluatiemoment. Observaties 2 en 4 daarentegen zijn onvolledig, en ongekend, want ongemeld op tijdstip τ. De uitdaging bestaat erin om de geobserveerde bovendriehoek te gebruiken om het aantal meldingen in de benedendriehoek te voorspellen.



Figuur 1: het zich voordoen ('occurrence') en het melden met vertraging ('reporting delay') van verschillende gebeurtenissen. Gebeurtenissen 2 en 4 zijn ongekend, want ongemeld op tijdstip τ. De gebroken lijnen geven de aggregatie weer van data in continue tijd naar discrete tijdsintervallen. De gebeurtenissen in continue tijd in de groene vlakken worden geaggregeerd tot 1 cel of observatie in de driehoek in figuur 2.

Terwijl figuur 1 start met continue tijdslijnen, laten de onderbroken lijnen in de figuur zien hoe observaties geaggregeerd kunnen worden tot een minder fijnmazig, discreet tijdsverloop, bijvoorbeeld door het aantal meldingen te aggregeren per maand, kwartaal of jaar. Een voorbeeld hiervan is een schadedriehoek met het aantal meldingen per ontstaans- en meldingsjaar, zoals geschetst in figuur 2. In deze driehoek staat de notatie N_{td} voor het aantal gebeurtenissen dat zich voordeed op (bijvoorbeeld dag, maand, kwartaal) t en gemeld werd met d vertraging (bijvoorbeeld in dagen, maanden, kwartalen). In ons literatuuronderzoek vonden we de representatie in figuur 2 niet alleen in actuariële bronnen, maar ook in epidemiologische toepassingen wordt de zogenaamde *reporting triangle* vaak gebruikt.

Occurrence period	Reporting delay: 0	…	$\tau - t$	…	$\tau - 1$
1	N_{10}	…	$N_{1,\tau-t}$	…	$N_{1,\tau-1}$
⋮					
t	N_{t0}	…	$N_{t,\tau-t}$		
⋮					
τ	$N_{\tau 0}$				

Figuur 2: voorstelling van geaggregeerde data als een driehoek. De gekleurde cellen tonen de aggregatie van de gebeurtenissen in continue tijd in figuur 1. Alleen de aantallen in de bovendriehoek zijn geobserveerd, terwijl de aantallen in de benedendriehoek voorspeld moeten worden.

EEN KLASSIEK VOORBEELD IS DE CHAINLADDERMETHODE DIE EEN MULTIPLICATIEVE SPECIFICATIE VOOR λ_{td} VOOROPSTELT

Een voor de hand liggende dimensie om de nowcastingliteratuur te structureren, bekijkt of het voorgesteld model voor continue dan wel discrete tijd geformuleerd is. Voor een overzicht van modellen in continue tijd, de bijhorende log-likelihood, en technieken om die te optimaliseren, verwijs ik graag naar ons artikel. Omdat gebeurtenissen veelal in tijdsintervallen geregistreerd worden, richt onze methodologische bijdrage zich op modellen die uitgaan van de discrete tijd representatie in figuur 2.

REGRESSIEMODELLEN IN DISCRETE TIJD: VAN KLASSIEKE CHAIN-LADDER TOT EEN NIEUWE, FLEXIBELE AANPAK

Een eerste aanpak die we vaak aantroffen in ons multidisciplinair literatuuronderzoek veronderstelt dat de N_{td} in de bovendriehoek in figuur 2 onafhankelijk en POI(λ_{td}) of NB(λ_{td}, φ) verdeeld zijn. De λ_{td} is hierbij de meldingsintensiteit. Een klassiek voorbeeld is de chain-laddermethode die een multiplicatieve specificatie voor λ_{td} vooropstelt, namelijk $\lambda_{td} = \lambda_t \,.\, p_d$ met als voorwaarde $p_0 + p_1 + \ldots + p_{\tau-1} = 1$, wat betekent dat alle claims worden gemeld met maximale vertraging gelijk aan τ-1 in figuur 2. Deze multiplicatieve structuur voor de meldingsintensiteit zorgt ervoor dat parameters redelijk eenvoudig geschat kunnen worden, zie ons artikel voor meer details. Merk daarbij op dat de chain-ladderspecificatie een stationair meldingsproces veronderstelt, waarbij de p_d onafhankelijk zijn van t.

Een tweede, alternatieve aanpak formuleert expliciet een model voor het zich voordoen van gebeurtenissen enerzijds en het meldingsproces anderzijds, een aanpak die meer in lijn ligt van de continue modellen en de tijdslijnen in figuur 1. Deze aanpak – die we grondig uitwerken in ons artikel – vertrekt van twee assumpties:

1 Het aantal gebeurtenissen N_t dat zich voordoet op t is POI verdeeld met intensiteit $\lambda_t = \exp(x_t^t \alpha)$ waarbij x_t een vector is van verklarende variabelen met betrekking tot tijdsperiode (bijvoorbeeld dag, maand, kwartaal) t.

2 Gegeven N_t, zijn de aantallen N_{td} multinomiaal verdeeld met kansen $p_{td} = p_{td}(\theta, x_{td})$. Deze meldingskansen sommeren tot 1 en bepalen zo een discrete kansverdeling, en maken gebruik van covariaten in x_{td}.

IN ONS ARTIKEL STELLEN WE EEN NIEUWE, SLIMME MANIER VOOR OM DE LIKELIHOOD TE OPTIMALISEREN: DE INZET VAN EEN EXPECTATION-MAXIMIZATION (EM) ALGORITME

De resulterende likelihood voor de geobserveerde gebeurtenissen kan geschreven worden als het product van een Poisson en een multinomiale likelihood. Echter, het optimaliseren van deze likelihood over de onbekende parameters is lastig omdat de likelihood niet gesplitst kan worden in een stuk dat alleen gebruikt maakt van de λ_t's enerzijds en de p_{td} kansen anderzijds. In ons artikel stellen we een nieuwe, slimme manier voor om de likelihood te optimaliseren: de inzet van een Expectation-Maximization (EM) algoritme. Het EM-algoritme vervangt (in de E-stap) het ontbrekende aantal gebeurtenissen in de benedendriehoek ('missing data') door hun verwachtingswaarde en optimaliseert vervolgens (in de M-stap) de parameters. Die laatste stap kan gebruik maken van standaard softwareroutines omdat de likelihood na invulling van de missing data wél gesplitst kan worden. Terwijl eerdere pogingen om het EM-algoritme in te zetten bij nowcastingproblemen zich beperkten tot eenvoudige specificaties voor het occurrenceproces in (1) en het meldingsproces in (2), zijn we nu in staat om nieuwe specificaties voor te stellen, die eerder te lastig waren om te schatten. Door covariaten in rekening te nemen, kunnen we flexibel tijds- en seizoenseffecten opnemen in het occurrenceproces of specifieke effecten inbouwen in het meldingsproces (bijvoorbeeld geen meldingen op officiële vakantiedagen of tijdens het weekend). Een casestudy in ons artikel focust op het modelleren van het aantal IBNR-schades aan de hand van verschillende nowcastingtechnieken. Hierbij laten we zien hoe empirische inzichten de specificatie van een passend occurrence en meldingsproces kunnen sturen. Bovendien benchmarken we de predicties aan de hand van onze voorgestelde aanpak met andere nowcastingmodellen uit de literatuur. Voordelen van onze aanpak zijn een verbeterde predictieve kracht en een verfijnd inzicht in het occurrence en meldingsproces enerzijds en de toekomstige meldingen anderzijds.

TOT SLOT

Nowcastingproblemen zijn relevant in verschillende disciplines, van actuariële wetenschappen tot epidemiologie. Vanuit onze actuariële kennis en discipline zijn we in staat om een generieke bijdrage te leveren die ook in andere toepassingen direct inzetbaar is. Tegelijkertijd reikt een grondige studie van de multidisciplinaire literatuur rond dit onderwerp ons relevante inzichten aan die ons uitdagen verder te denken dan het specifieke actuariële probleem. Een mooie win-win! ■

Katrien Antonio geeft het stokje door aan Jan Dhaene.

1 – Zie het artikel Modelling the occurrence of events subject to a reporting delay via an EM algorithm, te verschijnen in Statistical Science. Het artikel is in open access beschikbaar via arxiv.

2 – Prof. Gerda Claeskens van KU Leuven, Jonas Crevecoeur, PhD, postdoc aan KU Leuven en Roel Verbelen, PhD, statistisch consultant bij Finity Consulting in Sydney.