Wetenschappelijk Onderzoek- en Datacentrum

Cahier 2024-4

Voorspellen voor de justitiële ketens

Een verkenning van verschillende technieken

Kennisinstituut voor de rechtsstaat

Cahier 2024-4

Voorspellen voor de justitiële ketens

Een verkenning van verschillende technieken

D.E.G. Moolenaar
F. ter Braak
B. Tims
M.S. Bargh

Cahier
De reeks Cahier omvat de rapporten van onderzoek dat door en in opdracht van het Wetenschappelijk Onderzoek- en Datacentrum is verricht. Opname in de reeks betekent niet dat de inhoud van de rapporten het standpunt van de Minister van Justitie en Veiligheid weergeeft.

Inhoud

Afkortingen

ANN	Artificial neural network
AR	Autoregressive
ARIMA	Autoregressive integrated moving average
ARIMAX	Autoregressive integrated moving average with exogenous variables
AUC-ROC	Area under the curve – receiving operating characteristic
AUC-PR	Area under the curve – precision-recall
BLUE	Best linear unbiased estimator
CA	correspondentie-analyse
CJIB	Centraal Justitieel Incassobureau
COVID-19	Coronavirusdisease 2019
DGP	Datagenererend proces
DJI	Dienst Justitiële Inrichtingen
ECM	Error correction model
ETS	Exponential smoothing with error, trend and seasonality
FDA	Fisher's discriminant analysis
FPR	False positive rate
ISD	Inrichting voor stelselmatige daders
IV	instrumentele variabelen
Jukebox	Justitieketenmodel, box 1 en 2
KFCV	k-fold cross-validation
KNN	K-nearest neigbours
LASSO	Least absolute shrinkage and selection operator
LDA	Linear discriminant analysis
LOOCV	Leave-one-out cross-validation
MA	Moving average
MCA	meervoudige correspondentie-analyse
MinJenV	Ministerie van Justitie en Veiligheid
MLE	Maximum likelihood estimator
MSE	Mean squared error
OLS	Ordinary least squares
OM	Openbaar Ministerie
PaG	Parket-Generaal
PCA	Principal components analysis
PMJ	Prognosemodel Justitiële Ketens
PR	precision-recall
QDA	Quadratic discriminant analysis
RNN	Recurrent neural network
ROC	receiving operating characteristic
Rvdr	Raad voor de rechtspraak
SCP	Sociaal en Cultureel Planbureau
SSB	Sociaal-statistisch bestand
ST-AR	Space-time autoregressive
SVM	Support vector machine
SVR	Support vector regression
VECM	Vector error correction model
WODC	Wetenschappelijk Onderzoek- en Datacentrum
XAI	Explainable artificial intelligence
ZM	Zittende magistratuur

Samenvatting

Beleidsmakers willen graag meer inzicht in de (maatschappelijke) kosten van criminaliteit, rechtshandhaving en conflictbeslechting. Daarom is het belangrijk om inzicht te hebben in de toekomstige trends op dit gebied, zodat de best mogelijke beleidsmatige en financiële beslissingen kunnen worden genomen. Hiervoor kan gebruik worden gemaakt van prognosemodellen. Voor het beleidsterrein van Justitie is reeds enige tijd geleden het Prognosemodel Justiele Ketens (PMJ) ontwikkeld. In dit rapport wordt onderzocht in hoeverre het haalbaar en nuttig is om nieuwe ontwikkelingen op het gebied van data en algoritmen toe te passen in het PMJ.

Prognosemodel Justitiële Ketens

Momenteel worden prognoses over geregistreerde criminaliteit, verdachten en alles wat erop volgt, en conflictbeslechting gemaakt met het PMJ. Dit model omvat vrijwel de hele veiligheidsketen, waaronder opsporing, vervolging en berechting, straffen en maatregelen, gevangeniswezen, justitiële jeugdinrichtingen, reclassering, gesubsidieerde rechtsbijstand in strafzaken en slachtofferzorg. Daarnaast bevat het model ook de civiele rechtspraak, de bestuursrechtspraak, rechtsbijstand in civiele en bestuurszaken en vreemdelingenbewaring. Het PMJ gebruikt prognoses van demografische, maatschappelijke en economische achtergrondfactoren (ook wel exogene variabelen genoemd) om de instroom aan het begin van de keten te voorspellen zoals bijvoorbeeld geregistreerde criminaliteit. Deze prognose wordt vervolgens gebruikt om prognoses te maken van de instroom van zaken bij het Openbaar Ministerie (OM), en hiermee prognoses voor de instroom bij de rechtbanken en vervolgens prognoses voor de benodigde sanctiecapaciteit. Het PMJ is een combinatie van structurele modellen, stock-flow modellen en tijdreeksmodellen en omvat ongeveer 6.600 vergelijkingen. De parameters van het theoretische model worden geschat met behulp van regressieanalyse op jaargegevens. Het resultaat is een statistisch model waarmee jaarlijks prognoses worden gegenereerd, die dienen als onderbouwing voor een groot deel van de begroting van het ministerie van Justitie en Veiligheid (MinJenV).

Uit eerdere externe evaluaties blijkt dat het PMJ goed in elkaar zit en dat gebruikers geen behoefte hebben aan een radicaal ander model. Maar het huidige PMJ is ontworpen in een periode waarin de beschikbaarheid van microdata beperkt was en een aantal technieken vaak wel theoretisch bekend waren, maar niet konden worden geïmplementeerd vanwege beperkingen in computertechnologie. Daarom is het zinvol om nu te onderzoeken of er in de afgelopen jaren ontwikkelingen zijn geweest op het gebied van data en prognosetechnieken, die meer of andere inzichten kunnen bieden.

Afbakening

Het doel van het PMJ is het maken van ramingen van de toekomstige capaciteitsbehoefte van de justitiële ketens. Het PMJ-model gaat ervanuit dat de hele toekomstige capaciteitsbehoefte gefinancierd kan worden. Er worden dus geen budgetrestricties aan het PMJ toegevoegd. Ook houdt het PMJ-model rekening met keteneffecten. Beide aspecten (capaciteitsbehoefte en keteneffecten) komen overeen

met de wens geuit in het eindrapport van de parlementaire verkenning uit 2023 naar het functioneren van de strafrechtsketen. De ramingen betreffen uitsluitend aantallen. Het PMJ-model doet niets met prijzen. Het PMJ maakt ramingen van die items waarop justitiële organisaties worden gefinancierd. Welke items dat precies zijn, verschilt per organisatie en wordt door de organisaties zelf in samenspraak met het ministerie van Justitie en Veiligheid bepaald en niet door het PMJ-model. Het PMJ-model is hierin dus volgend en niet leidend. Het PMJ bepaalt niet wat en hoe er gefinancierd moet worden, alleen maar hoeveel er gefinancierd moet worden, gegeven wat en hoe. Als de wijze van financiering wordt aangepast, wordt het PMJ-model aangepast. Voorwaarde voor het PMJ-model is wel dat de criteria waarop wordt gefinancierd, kwantificeerbaar en meetbaar zijn. Omdat de wijze van financiering een beslissing is die buiten het PMJ om wordt genomen, zal dit rapport hier verder niet op ingaan. Ook de ramingen zelf en de trefzekerheid ervan worden hier niet besproken. Deze zijn terug te vinden in andere rapporten van het Wetenschappelijk Onderzoek- en Datacentrum. De focus van dit rapport ligt op modellen waarmee trends voor (lange-termijn) strategische doeleinden kunnen worden voorspeld en niet op voorspelmodellen voor operationele of forensische doeleinden, zoals 'predictive policing' of 'predictive sentencing'.

Aanleiding, onderzoeksvraag en randvoorwaarden

Tijdens de begrotingsbehandeling 2019 heeft de Minister voor Rechtsbescherming de toezegging gedaan dat hij bereid is nog eens naar het PMJ-model te kijken. Naar aanleiding hiervan is besloten tot een 2-sporen aanpak om het huidige PMJ te herzien. Spoor 1 betreft onderhoud van en kleine verbeteringen en aanvullingen op het huidige PMJ. Spoor 2 betreft het fundamenteel onderzoeken van methoden en technieken voor betere ramingen. Spoor 2 is opgedeeld in drie fases. In de eerste fase is een inventarisatie gemaakt van de behoefte van de eindgebruikers van PMJ. Deze fase is inmiddels afgerond in 2020. In de tweede fase is gekeken in hoeverre nieuwe ontwikkelingen op het gebied van data en technieken benut zouden kunnen worden in het PMJ. In de derde fase zullen enkele veelbelovende technieken in de vorm van pilots nader worden uitgewerkt.
Dit rapport doet verslag van de tweede fase. Daarin zijn een groot aantal technieken bekeken die in potentie relevant kunnen zijn voor het PMJ. Dat wil zeggen dat met deze selectie van technieken in principe het doel van het PMJ bereikt zou kunnen worden, namelijk het maken van ramingen van de capaciteitsbehoefte van de justitiële ketens prognoses ten behoeve van begroting. Technieken die niet geschikt zijn voor dit doel, zijn buiten beschouwing gelaten. De technieken zijn beoordeeld op onderstaande aspecten:

1 Uitlegbaarheid van het algoritme. Hoe makkelijk is het om in eenvoudige termen uit te leggen wat het algoritme doet? Kortom, hoe intuïtief is het algoritme?
2 Eenvoud van het algoritme. Hoe simpel is het algoritme vanuit een wiskundig/statistisch standpunt?
3 Implementeerbaarheid. Hoeveel werk kost het om het algoritme te implementeren?
4 Domeinkennis. Is het mogelijk om domeinkennis in te brengen in het algoritme?
5 Ketenconsistentie. Is het mogelijk om met een algoritme tot een ketenconsistent model te komen? Dat wil zeggen een model waarbij de uitstroom van de ene partner de instroom voor een volgende partner vormt?
6 Tijdscomponent. Is het mogelijk om een tijdscomponent in het algoritme mee te nemen? Dat wil zeggen kan het algoritme dynamisch een prognose maken voor de

(middel)lange termijn of moeten sommige onderdelen constant worden verondersteld?
7 Omgang met ruis in de data. Kan het algoritme omgaan met ruis in de data of moet de kwaliteit van de data heel hoog zijn?
8 Privacy. In hoeverre zijn micro-data nodig of is het algoritme ook toepasbaar op geaggregeerde data? En als voor micro-data wordt gekozen, kunnen de resultaten dusdanig worden geaggregeerd, dat ze geschikt zijn voor verdere verwerking in een geaggregeerd model?
9 Rekentijd. Hoeveel rekentijd kost het om tot prognoses te komen?
10 Inhoudelijke uitlegbaarheid van de prognoses. Zijn de prognoses logisch en in eenvoudige termen uit te leggen? Zijn de prognoses naar specifieke input variabelen herleidbaar of hangt alles met alles samen?
11 Rechtvaardigheid ('fairness'). In welke mate kunnen ongelukkige keuzes of beslisregels ertoe leiden dat het algoritme onbedoeld een discriminerend karakter krijgt?

Er zijn heel veel technieken om ramingen te maken. Om bruikbaar te zijn voor het PMJ, moeten de technieken aan een aantal randvoorwaarden voldoen, die vooral voortkomen uit het begrotingsproces en de wensen van de eindgebruikers van de PMJ-ramingen:

- Het model moet ketenconsistent zijn. De prognose van de uitstroom van de ene ketenpartner moet doorwerken in de prognose van de instroom van de daaropvolgende ketenpartner.
- Er moet zeven jaar vooruit voorspeld kunnen worden, d.w.z. de begrotingshorizon (vijf jaar) plus de twee jaren tussen het laatst bekende realisatiejaar en het eerste begrotingsjaar.
- De prognoses moeten inhoudelijk uitlegbaar zijn. Beleidsmakers willen graag kunnen begrijpen waarom de prognoses zijn zoals ze zijn. In de praktijk betekent dit dat ze herleidbaar moeten zijn naar concrete inputvariabelen en dat de geschatte relatie een zekere mate van logica moet bevatten.
- Vanwege de planning van het begrotingsproces moeten de parameters van het model jaarlijks half november geactualiseerd zijn. Omdat een aantal gegevens pas eind september beschikbaar zijns, betekent dit in de praktijk dat de actualisering binnen een periode van circa zes weken moet plaatsvinden.
- Het gekozen algoritme moet rechtvaardig zijn. Gemaakte keuzes of beslisregels mogen er niet onbedoeld toe leiden dat het algoritme een discriminerend karakter krijgt.

Alternatieve methoden

De technieken die bekeken zijn afkomstig uit de machine learning en de econometrie. Hoewel er grote overlap is tussen de technieken die in de econometrie en in machine learning worden gebruikt, werkt de econometrie meer vanuit de theorie en is machine learning meer datagedreven. Grofweg vallen de alternatieve methoden uiteen in vier categorieën: een aanscherping van de wijze waarop de parameters van het huidige PMJ-model worden geschat, een andere specificatie van (delen van) het model, methoden die betrekking hebben op een grotere benutting van de dataset waarmee de modellen worden geschat en getest en het combineren van methoden of steekproeven.

Aanscherping van het huidige PMJ

Lineaire regressie is een klasse van eenvoudige en daarom veelvuldig toegepaste algoritmes op lineaire of gelineariseerde modellen. Tot nu toe wordt in het PMJ vooral gewone kleinste kwadraten toegepast. Een ander lineair algoritme dat mogelijk interessant kan zijn voor het PMJ is lineaire regressie met elastic net regularisatie en in het bijzonder ridge regressie. Bij deze methode wordt een straf gezet op een te grote complexiteit van het model, dat wil zeggen te veel exogene variabelen in het model. Indien de data problemen bevat, zoals bijvoorbeeld uitschieters, slecht gemeten of onderling gecorreleerde exogene variabelen, niet waargenomen categorieën en/of kleine of scheef verdeelde steekproeven, dan kan afhankelijk van het type probleem voor een aangepast lineair algoritme worden gekozen. Maar vaak kunnen deze problemen ook op een andere, eenvoudigere wijze kunnen worden opgelost.

Andere specificatie van (delen van) het PMJ

Als er aantallen moeten worden voorspeld, kan zowel lineaire als niet-lineaire regressie worden toegepast, zoals bijvoorbeeld lineaire of niet-lineaire tijdreeksanalyse of survivalanalyse (niet-lineair). Tijdreeksanalyse wordt veelvuldig gebruikt voor het maken van prognoses, omdat deze methode relatief makkelijk te implementeren is en in veel softwarepakketten standaard geïntegreerd is. In het huidige PMJ wordt reeds in beperkte mate lineaire tijdreeksanalyse toegepast. Wel moet de data aan een aantal voorwaarden voldoen. Een nadeel is dat tijdreeksanalyses vooral geschikt zijn voor korte-termijnprognoses, omdat tijdreeksmodellen op de lange termijn de neiging hebben terug te keren naar het gemiddelde van het proces. Andere vormen van niet-lineaire regressie die mogelijk interessant kunnen zijn voor PMJ zijn algoritmes voor duurgegevens. Met name een (semi-)parametrische survivalanalyse van duurgegevens lijkt veelbelovend voor die onderdelen van het justitieveld waarvoor de duur niet vooraf bekend is, zoals de tbs- of pij-maatregel of voorlopige hechtenis.

Voor de prognoses van bepaalde keuzemomenten zijn er algoritmes zoals discriminantenanalyse, naïeve Bayes algoritme of logistische regressie. Voor het PMJ lijkt logistische regressie een logische optie: binnen de justitiële ketens zijn veel keuzemomenten. Moet een verdachte preventief worden gehecht? Moet een verdachte wel of niet worden vervolgd en/of berecht? Welke type sanctie moet worden opgelegd? Dit zijn typisch keuzes die met een logistische regressie kunnen worden voorspeld. Desalniettemin zijn er ook beperkingen. Het doel van het PMJ is om zeven jaar vooruit te voorspellen. Het is lastig om het tijdsaspect in een logistische regressie op te nemen. Bovendien wordt logistische regressie toegepast op microdata. De vraag is of de kans op schending van de privacy opweegt tegen een vergroot inzicht in de prognoses. Discriminantenanalyse is een eenvoudig en rekenkundig efficiënt algoritme, maar vooral geschikt voor kleine steekproeven met een beperkt aantal exogene variabelen. Ook het naïeve Bayes algoritme is technisch gezien niet moeilijk, maar de resultaten zijn minder intuïtief als men geen kennis van of affiniteit met kansverdelingen heeft. Dit maakt beide algoritmes minder bruikbaar voor het PMJ.

Er zijn ook algoritmes waarmee zowel aantallen als keuzes kunnen worden voorspeld, zoals, k-nearest neighbours, beslisbomen, support vector machines en neurale netwerken. Deze non-parametrische algoritmes hebben als voordeel dat er vooraf weinig aannames over de data worden gedaan. Verder verschillen de voor- en nadelen per algoritme. Alles overziend zouden k-nearest neighbours en beslisbomen een goede aanvulling voor het PMJ kunnen zijn, waarbij belangrijkste nadelen van respectievelijk

het niet kunnen herleiden naar specifieke achtergrondfactoren dan wel het ontbreken van de tijdscomponent zullen moeten worden afgewogen tegen de voordelen. Support vector machines en neurale netwerken vallen af voornamelijk omdat de rekentijd van deze algoritmes groot is en de actualisering van de modellen niet binnen de beschikbare tijd (circa zes weken) op verantwoorde wijze kan worden verwezenlijkt. Een ander nadeel dat met name bij neurale netwerken speelt, is dat alles met elkaar samenhangt en prognoses niet inhoudelijk uitlegbaar zijn.

Grotere benutting van de dataset

Er zijn meerdere methoden om de huidige steekproef ruimer te benutten. Een manier om de steekproef breder te benutten is om de validatiefouten of de standaardfouten en de betrouwbaarheidsintervallen van de geschatte parameters te verbeteren. Door middel van bootstrapping of kruisvalidatie kunnen uit dezelfde steekproef meerdere subsets worden getrokken waarmee iedere keer hetzelfde model wordt geschat. Deze methoden worden gebruikt om verschillende theoretische modellen en/of algoritmes met elkaar te vergelijken op het punt van voorspelkracht. Bootstrapping kan vooral in combinatie met andere methoden interessant zijn voor het PMJ. Het nadeel is dat bij een groot aantal modellen, algoritmes en/of waarnemingen kan dit zeer rekenintensief zijn. Bovendien zijn beide algoritmes wat moeilijker toepasbaar op tijdreeksen. Mede omdat het PMJ doorgaans niet uitsluitend wordt beoordeeld op voorspelkracht, maar op meerdere maatstaven, waaronder theoretische juistheid en inhoudelijke uitlegbaarheid, lijkt met name bij kruisvalidatie het marginale voordeel van betere voorspelmaatstaven en kennis over de verdeling van de parameters niet op te wegen tegen het nadeel van de grotere rekenintensiteit.

Een andere manier om de steekproef breder te benutten is om in plaats van individuele exogene variabelen combinaties van exogene variabelen in het model worden opgenomen, bijvoorbeeld door middel van principale componenten analyse of factoranalyse. Het grote nadeel is dat de resulterende prognoses niet goed herleidbaar zijn naar specifieke achtergrondfactoren en dus niet inhoudelijk uitlegbaar zijn. Daarom is de toepassing van principale componenten analyse of factoranalyse niet wenselijk voor PMJ.

Combineren van methoden of steekproeven

Er zijn een aantal manieren waarop verschillende prognoses van dezelfde variabele kunnen worden gecombineerd. Bij bagging en boosting worden prognoses van hetzelfde algoritme maar verschillende steekproeven gecombineerd. Bagging, ofwel bootstrap aggregating, zou voor het PMJ interessant kunnen zijn, omdat het relatief eenvoudig te implementeren is en het model zelf ook niet al te complex wordt, zodat de resultaten nog steeds uitlegbaar zijn. Bagging is tevens een essentieel onderdeel van het random forest algoritme. Random forest is een woud van beslisbomen, waarbij elke beslisboom op een andere gebootstrapte steekproef van waarnemingen en/of een andere willekeurige selectie van de exogene variabelen wordt opgebouwd. Boosting is minder interessant omdat het zeer hoge kwaliteit van de data vereist zonder uitschieters en ruis en het is rekenintensief, waardoor de prognoses niet binnen de beschikbare tijd kunnen worden geactualiseerd.

Bij ensemble averaging en stacking kunnen prognoses van modellen geschat met verschillende algoritmes op dezelfde steekproef worden gecombineerd. Bij ensemble averaging wordt een gemiddelde berekend, terwijl bij stacking een metamodel wordt

geformuleerd. Ensemble averaging is een veelbelovende techniek voor het PMJ, omdat het relatief eenvoudig te implementeren is en er sowieso in de testfase vaak al verschillende algoritmes worden uitgeprobeerd. Tot nu toe werd op basis van verschillende criteria uiteindelijk één model gekozen, hoewel de onderlinge verschillen in voorspelkwaliteit vaak gering waren. Als alternatief zou het gemiddelde van de prognoses van meerdere modellen kunnen worden berekend. Zo zou bijvoorbeeld de instroom bij het OM kunnen worden voorspeld op basis van het aantal verdachten dat door de politie wordt geregistreerd, maar ook door middel van een tijdreeksanalyse. De uiteindelijke prognose wordt dan het gemiddelde van beide algoritmes. Stacking is ook een optie zolang het metamodel niet te ingewikkeld is. Maar het nadeel is dat het veel rekentijd kost en dat er (relatief) veel data nodig is.

Conclusie en aanbevelingen

Machine learning modellen zijn vooral datagedreven en dus vooral gebaseerd op correlaties en niet zozeer op causale verbanden, dit in tegenstelling tot econometrische modellen. Een recente ontwikkeling is de groeiende aandacht voor causale en uitlegbare machine learning technieken, de zogenoemde 'explainable artificial intelligence' (XAI). Daarmee groeien de machine learning modellen en de econometrische modellen naar elkaar toe. Dit lijkt ook de meest belovende ontwikkelingsrichting voor het PMJ te zijn, maar XAI is op dit moment nog wel een kennisgebied in ontwikkeling.

Gegeven de aard van de data, het doel van PMJ en de randvoorwaarden, zijn de meest belovende alternatieve algoritmes:

- algoritme waarbij een straf wordt gezet op een te grote complexiteit van het model (lineaire regressie met elastic net regularisatie);
- analyse van duurgegevens tbs- of pij-maatregelen of voorlopige hechtenis (survivalanalyse);
- algoritme dat ervanuit gaat dat vergelijkbare kenmerken van de achtergrondfactoren leidt tot vergelijkbare waarden van de te voorspellen variabele (k-nearest neighbours);
- algoritme voor keuzemomenten, zoals het type straf wat moet worden opgelegd (logistische regressie, beslisboom);
- grotere benutting van de bestaande dataset door meerdere steekproeven uit de dezelfde dataset te trekken (bagging van gebootstrapte steekproeven);
- combinatie van meerdere beslisbomen door middel van bagging (random forest);
- combinatie van de uitkomsten van verschillende algoritmes (ensemble averaging);
- tijdreeksanalyse ten behoeve van ensemble averaging.

Grotendeels voldoen deze algoritmes aan de randvoorwaarden, maar soms zullen er concessies moeten worden gedaan. In een vervolgonderzoek zullen een aantal pilots met deze algoritmes worden uitgevoerd om te kijken of deze algoritmes ook daadwerkelijk tot een hogere voorspelkwaliteit leiden.

1 Inleiding

Beleidsmakers willen graag meer inzicht in de ontwikkeling van de criminaliteit, rechtshandhaving en conflictbeslechting en in de kosten hiervan voor de samenleving. Daarom is het belangrijk om inzicht te hebben in de toekomstige trends op dit gebied, om zo de best mogelijke beleidsmatige en financiële beslissingen te kunnen nemen. Hiervoor kan gebruik worden gemaakt van prognosemodellen. Voor het beleidsterrein van Justitie is reeds enige tijd geleden het Prognosemodel Justiele Ketens (PMJ) ontwikkeld. In dit rapport wordt onderzocht in hoeverre het haalbaar en nuttig is om nieuwe ontwikkelingen op het gebied van data en algoritmen toe te passen in het PMJ.

1.1 Geschiedenis van het Prognosemodel Justitiële Ketens

Van 1989 tot en met 1996 maakte het ministerie van Justitie grofweg eens in de vier jaar een raming van de capaciteitsbehoefte in de justitiële inrichtingen. Deze prognoses werden gemaakt onder de verantwoordelijkheid van departementale projectgroepen. In de tussenliggende jaren werden soms op ad hoc-basis en per sector prognoses gemaakt. In 1996 kwam er kritiek van de Nationale Ombudsman, n.a.v. klachten van tbs-passanten, op de methodiek voor de tbs-prognoses. In 1997 leverde de Algemene Rekenkamer, n.a.v. heenzendingen en de tenuitvoerlegging van sancties, kritiek op de prognoses voor het gevangeniswezen, het proces van besluitvorming tussen kerndepartement, Dienst Justitiële Inrichtingen (DJI) en het Openbaar Ministerie (OM) en wees op het gebrek aan feitelijke, actuele en betrouwbare gegevens (Van der Heijden, 2002). Daarom zegde de toenmalige Minister van Justitie aan de Tweede Kamer toe dat het ministerie voortaan elk jaar een raming van de behoefte aan sanctiecapaciteit zou maken.

In 1998 werden de eerste bouwstenen gelegd voor het prognosemodel voor de Nederlandse (straf)rechtspraak (Van der Torre & Van Tulder, 2001). Het oorspronkelijke model gebruikte demografische en economische prognoses om de geregistreerde criminaliteit te voorspellen, die op hun beurt werden gebruikt om prognoses te maken van de vereiste sanctiecapaciteit in het gevangeniswezen en voor taakstraffen voor meerderjarigen. Later werd het model verfijnd en uitgebreid met een jeugdsector (Huijbregts et al., 2001). In 2003 hebben de ministeries van Justitie, Financiën, Binnenlandse Zaken en Koninkrijksrelaties en Algemene Zaken de afspraak gemaakt dat het ministerie van Justitie een integraal model voor de justitiële ketens zou ontwikkelen. Besloten werd om het bestaande model voor de sanctiecapaciteit uit te breiden tot het PMJ. Het huidige PMJ omvat vrijwel de hele veiligheidsketen, waaronder opsporing, vervolging en berechting, straffen en maatregelen, vreemdelingenbewaring, gevangeniswezen, justitiële jeugdinrichtingen, reclassering, gesubsidieerde rechtsbijstand in strafzaken en slachtofferzorg (zie Moolenaar et al., 2004). Ook de civiele rechtspraak, de bestuursrechtspraak (Leertouwer et al., 2005) en rechtsbijstand in civiele en bestuurszaken (Leertouwer et al., 2007) zitten in het PMJ. De vreemdelingenketen maakt geen onderdeel uit van het PMJ, met uitzondering van vreemdelingenbewaring en gesubsidieerde rechtsbijstand in vreemdelingenzaken.

1.2 Aanleiding voor dit onderzoek

Tijdens de begrotingsbehandeling 2019 (Handelingen II 2018/19, 25, item 29, p. 62) heeft de Tweede Kamer naar aanleiding van een discussie over de financiering van de rechtspraak en het OM aan de Minister voor Rechtsbescherming gevraagd wat hij gaat doen om het PMJ-model te verbeteren en het op een betere slimmere manier te gebruiken. Hierop heeft de minister geantwoord (Handelingen II, 2018/19, 27, item 10, p. 6) dat hij bereid is nog eens naar het PMJ-model te laten kijken en daar waar nodig te finetunen.

Recentelijk heeft de vaste Kamercommissie voor Justitie en Veiligheid een parlementaire verkenning uitgevoerd naar de prestaties in de strafrechtketen (*Kamerstukken II* 2022/23, bijlage bij 29 279, nr. 804). De algemene conclusie is dat er meer op de strafrechtketen als geheel moet worden gestuurd. De coördinatie van de keten is gebrekkig en een integrale aanpak ontbreekt. Ook is er onvoldoende (juiste) informatie om te kunnen sturen op de uitvoering in de keten. Het kost te veel tijd om de informatie te verwerken, de informatiesystemen zijn ontoereikend en de kwaliteit van de data is onvoldoende. Tot slot ontbreekt het aan ketenperspectief in de financiering. Als er onvoldoende geld is, concentreren organisaties zich op de eigen hoofdtaken en hebben samenwerking en gezamenlijke doelen geen prioriteit. Bij het begroten is niet leidend wat er nodig is in de keten maar het beschikbare budget. Eén van de aanbevelingen die de rapporteurs geven is om "alternatieven voor de doorslaggevende rol van het Prognose Model Justitiële Ketens te ontwikkelen". Ten aanzien van de financiering en het PMJ heeft de Minister voor Rechtsbescherming geantwoord dat recent onderzoek (*Kamerstukken I* 2020/21, 35 300-VI, nr BK) heeft geconcludeerd dat er geen aanleiding is om de bekostigingssystemen van de organisaties als zodanig aan te passen, dat PMJ een belangrijk hulpmiddel is om zicht te houden op de te verwachten werklast voor de justitiële ketens en de ontwikkelingen daarin en dat het PMJ maar één aspect van de bekostiging van de strafrechtsketen is (*Kamerstukken II*, 2023/24, 29 279, nr. 836).

Reeds naar aanleiding van de begrotingsbehandeling 2019 heeft de bestuursraad van het ministerie van Justitie en Veiligheid (MinJenV) eind 2018 ingestemd met een 2-sporen aanpak om het huidige PMJ te herzien. Spoor 1 betreft onderhoud van en kleine verbeteringen en aanvullingen op het huidige PMJ. Spoor 2 betreft het fundamenteel onderzoeken van methoden en technieken voor betere ramingen. Spoor 2 is opgedeeld in 3 fases. In fase 1 is een inventarisatie gemaakt van de behoefte van de eindgebruikers van PMJ. Deze fase is inmiddels afgerond (zie De Poot et al., 2020). In de tweede fase zal worden gekeken in hoeverre nieuwe ontwikkelingen op het gebied van data en technieken benut zouden kunnen worden in het PMJ. Dit rapport doet verslag van deze verkenning. In de derde fase zullen enkele veelbelovende technieken in de vorm van een pilots nader worden uitgewerkt.

1.3 Afbakening

Allereerst is het belangrijk om het PMJ-model te onderscheiden van het PMJ-proces. Het PMJ-model is een verzameling van databases, wiskundig-statistische vergelijkingen en algoritmes die door het gebruik van algoritmes en databases worden geschat wat uiteindelijk leidt tot een groot aantal ramingen op justitieterrein. Het PMJ-model valt onder de verantwoordelijkheid van het Wetenschappelijk Onderzoek- en Datacentrum (WODC). Het PMJ-proces gaat over het gebruik van de ramingen van het

PMJ-model, zoals bijvoorbeeld het toevoegen van beleidseffecten, de vertaling naar budget en de communicatie met het ministerie van Financiën hierover. Dit valt onder de verantwoordelijkheid van directie Financieel-Economische Zaken van het ministerie van Justitie en Veiligheid.

1.3.1 PMJ-model

Het doel van het PMJ-model is het maken van ramingen van de capaciteitsbehoefte van de justitiële ketens. Het PMJ-model gaat ervanuit dat de gehele capaciteitsbehoefte gefinancierd kan worden. Er zitten dus geen budgetrestricties in het PMJ. De ramingen betreffen uitsluitend aantallen. Het PMJ-model doet niets met prijzen. Er worden ramingen gemaakt van die items waarop justitiële organisaties worden gefinancierd. Wat dat precies is, verschilt per organisatie en wordt door de organisaties zelf in samenspraak met ministerie van Justitie en Veiligheid bepaald en niet door het PMJ-model. Het PMJ-model is hierin dus volgend en niet leidend. Het PMJ bepaalt niet wat en hoe er gefinancierd moet worden, alleen maar hoeveel er gefinancierd moet worden, gegeven wat en hoe.

De ramingen gemaakt met het PMJ-model zijn beleidsneutraal. Dat wil zeggen dat ontwikkelingen en beleid uit het verleden worden doorgetrokken naar de toekomst. Het effect van reeds vastgesteld beleid is alleen verdisconteerd in de prognoses voor zover de invloed ervan reeds in de gebruikte gegevens zichtbaar is. Ontwikkelingen als gevolg van inmiddels ingevoerd of voorgenomen beleid of wetgeving, die in verband met een recente invoeringsdatum nog niet in de statistieken (kunnen) zijn verwerkt, blijven dan ook buiten beschouwing in de beleidsneutrale ramingen.

De beleidsneutrale ramingen zijn momentopnames. Ze zijn gebaseerd op de beschikbare kennis op het moment van berekenen. Zowel de gelegde relaties in het PMJ als de van elders betrokken ramingen van de achtergrondfactoren brengen onzekerheden met zich mee. De beleidsneutrale ramingen moeten derhalve niet als een vaststaand gegeven worden beschouwd, maar voornamelijk als een signaal. Ze geven aan wat er zou kunnen gebeuren indien er niets verandert. Door onverwachte gebeurtenissen, onbekende factoren, meetfouten, misspecificatie van het model en niet in de raming verwerkte wijzigingen in wet- en regelgeving en beleid kunnen de werkelijke ontwikkelingen afwijken van de ramingen.

1.3.2 PMJ-proces

Aanvulling van de beleidsneutrale raming met ontwikkelingen als gevolg van voorgenomen beleid of wetgeving wordt door de beleidsdirecties van het MinJenV, het Parket-Generaal (PaG), de Raad voor de rechtspraak (Rvdr) en enkele uitvoeringsorganisaties verzorgd. Deze organisaties inventariseren welke ontwikkelingen nog niet (volledig) zijn verwerkt in de beleidsneutrale ramingen en wat de consequenties daarvan kunnen zijn voor de justitiële ketens. Het gaat hierbij om nog niet (volledig) zichtbare effecten van reeds in gang gezet beleid en in beperkte mate om effecten van nieuw voorgenomen, reeds goedgekeurd beleid. Om deze effecten in te schatten worden vaak expert opinions of simulatiemodellen gebruikt (Smit & Choenni, 2014). De ramingen van deze zogenoemde beleidseffecten vallen onder de verantwoordelijkheid van de beleidsdirecties, PaG, Rvdr en de uitvoeringsorganisaties zelf. Soms wordt het PMJ-model gebruikt als simulatiemodel.

De beleidsneutrale ramingen vormen samen met de door de beleidsdirecties, PaG, Rvdr en de uitvoeringsorganisaties geraamde beleidseffecten de beleidsrijke ramingen. De beleidsrijke ramingen worden door de directie Financieel-Economische Zaken van het ministerie van Justitie en Veiligheid vertaald naar een financiële behoefte. Dit dient als input voor het onderhandelingstraject met het ministerie van Financiën. Het resultaat hiervan komt uiteindelijk in de jaarlijkse begroting te staan. De begroting geeft dus niet noodzakelijkerwijs de behoefte weer maar alleen dat deel van de behoefte waarvoor geld beschikbaar is.

1.3.3 Focus van het onderzoek

Afgaande op het rapport van de parlementaire verkenning (*Kamerstukken II* 2022/23, bijlage bij 29 279, nr. 804) en de eerdere begrotingsbehandeling 2019 (Handelingen II 2018/19, 25, item 29, p. 62 en 27, item 10, p. 6) lijkt de Tweede Kamer vooral een probleem te hebben met de wijze waarop een aantal justitiële organisaties worden gefinancierd (voornamelijk prijs x productieaantallen).
Dit wordt echter niet door het PMJ-model of het PMJ-proces bepaald, maar door de justitiële organisaties zelf in samenspraak met het moederdepartement. Het PMJ-model en het PMJ-proces zijn daarin slechts volgend en niet leidend. Als de wijze van financiering wordt aangepast, dan kan het PMJ-model hierop worden aangepast. Voorwaarde voor het PMJ-model is wel dat de criteria waarop wordt gefinancierd, kwantificeerbaar en meetbaar zijn. Omdat de wijze van financiering een beslissing is die buiten het PMJ-proces om wordt genomen, zal dit rapport hier verder niet op ingaan.

Uit de tekst van het eindrapport van de parlementaire verkenning wordt niet duidelijk waarop de aanbeveling ten aanzien van het PMJ-model is gebaseerd. Het PMJ-model maakt namelijk ramingen van de behoefte en niet van de productie en het PMJ-model houdt rekening met keteneffecten, wat overeenkomt met de wens geuit in het eindrapport van de parlementaire verkenning. Daarom zal dit rapport ook niet nader ingaan op de uitkomsten van de parlementaire verkenning. Omdat het PMJ-proces niet onder de verantwoordelijkheid van het WODC valt, zal het PMJ-proces ook niet verder besproken worden in dit rapport. Ook de ramingen zelf worden hier niet besproken. De meest recente beleidsneutrale ramingen zijn te vinden in Tims et al. (2023). De meest recente analyse van de nauwkeurigheid van de PMJ-ramingen is te vinden in Moolenaar et al. (2021). In het vervolg van dit rapport zal het PMJ-model kortweg worden aangeduid als *het PMJ*.

De focus van dit rapport ligt op modellen waarmee de trends voor (lange-termijn) strategische doeleinden kunnen worden voorspeld en niet op voorspelmodellen voor operationele of forensische doeleinden. Voorbeelden van voorspellen voor operationele doeleinden zijn ‘predictive policing’, waarbij hotspots van criminaliteit worden geïdentificeerd zodat de politie weet waar ze moet patrouilleren, of ‘predictive sentencing’, waarbij de meest geschikte straf wordt voorspeld op basis van de kenmerken en omstandigheden van de verdachte. Dat neemt niet weg dat sommige methoden die voor strategische doeleinden worden gebruikt, ook kunnen worden gebruikt voor operationele doeleinden. Ook scenario-analyses en simulaties vallen buiten het bestek van dit rapport, omdat deze methoden zich meer bezighouden met complexe 'wat als’-vragen. Met andere woorden, in dit rapport concentreren we ons op basismethoden, die kunnen worden gezien als componenten van andere, meer complexe methoden.

1.4 Onderzoeksvragen

Het doel van het PMJ is om (middel)lange-termijn geaggregeerde ramingen te maken ter onderbouwing voor een groot deel van de begroting van het ministerie van Justitie en Veiligheid. Dit rapport geeft antwoord op de vraag in hoeverre nieuwe ontwikkelingen op het gebied van data en algoritmen benut zouden kunnen worden in het PMJ voor bijvoorbeeld het voorspellen van de geregistreerde criminaliteit, het aantal verdachten dat wordt vervolgd en berecht of het aantal zaken waarin conflicten worden beslecht. Dit zou kunnen leiden tot implementatie van nieuwe (sub)modellen en/of nieuwe algoritmen in (delen van) het PMJ.

In de praktijk worden de begrippen algoritme en model vaak door elkaar gebruikt, omdat het verschil niet voor iedereen duidelijk is. Daarbij komt nog dat het begrip model, afhankelijk van de wetenschappelijke discipline, verschillende betekenissen heeft. Het varieert van een volwaardig hypothetisch, deductief systeem tot een verzameling wetten of regels met betrekking tot een systeem. Daarom worden in het vervolg onderstaande begrippen gehanteerd.

- Het theoretisch model geeft de structuur van de causale relatie weer tussen twee of meerdere variabelen.
- Het algoritme is de wijze waarop de parameters van het theoretische model worden geschat.
- Het empirisch model is het resultaat van de toepassing van het algoritme op de data. Het empirische model wordt gebruikt on een voorspellingen te maken op basis van nieuwe datapunten.

Indien een onderwerp zowel op het theoretisch model als het empirisch model van toepassing is, spreken we gewoon van een model. Dit rapport richt zich voornamelijk op alternatieve algoritmen, maar een alternatief algoritme kan gepaard gaan met een alternatieve specificatie van het model. Bij de zoektocht naar alternatieve algoritmen is met name gelet op onderstaande aspecten:

1 *Inhoudelijke uitlegbaarheid van het algoritme.* Hoe moeilijk of makkelijk is het om in eenvoudige termen uit te leggen wat het algoritme doet? Kortom, hoe intuïtief is het algoritme?
2 *Eenvoud van het algoritme.* Hoe simpel of complex is het algoritme vanuit een wiskundig/statistisch standpunt?
3 *Implementeerbaarheid.* Hoeveel werk kost het om het algoritme te implementeren?
4 *Domeinkennis.* Is het mogelijk om domeinkennis in te brengen in het algoritme?
5 *Ketenconsistentie.* Is het mogelijk om met een algoritme tot een ketenconsistent model te komen, dat wil zeggen een model waarbij de uitstroom van de ene partner de instroom voor een volgende partner vormt?
6 *Tijdscomponent.* Is het mogelijk om een tijdscomponent in het algoritme mee te nemen? Dat wil zeggen kan het algoritme dynamisch een prognose maken voor de (middel)lange termijn of moeten sommige onderdelen constant worden verondersteld?
7 *Omgang met ruis in de data.* Kan het algoritme omgaan met ruis in de data of moet de kwaliteit van de data heel hoog zijn?
8 *Privacy.* In hoeverre zijn micro-data nodig of is het algoritme ook toepasbaar op geaggregeerde data? En als voor micro-data wordt gekozen, kunnen de resultaten dusdanig worden geaggregeerd, dat ze geschikt zijn voor verdere verwerking in het model?

9 *Rekentijd.* Hoeveel rekentijd kost het om tot prognoses te komen?
10 *Inhoudelijke uitlegbaarheid van de prognoses.* Zijn de prognoses logisch en in eenvoudige termen uit te leggen? Zijn de prognoses naar specifieke input variabelen herleidbaar of hangt alles met elkaar samen?
11 *Rechtvaardigheid ('fairness').* In welke mate kunnen ongelukkige keuzes of beslisregels ertoe leiden dat het algoritme onbedoeld een discriminerend karakter krijgt?

Merk op dat in essentie het doel van alle algoritmes die in dit rapport worden behandeld, is om de data op basis van bepaalde kenmerken in verschillende categorieën onder te verdelen zonder aan deze categorieën een waardeoordeel te verbinden. Door een ongelukkige keuze van kenmerken, kunnen deze categorieën een discriminerend karakter krijgen. Hierdoor kan het algoritme (indirect) ten onrechte bepaalde bevolkingsgroepen benadelen op basis van bepaalde persoonskenmerken. De laatste jaren is hiervoor veel aandacht. Een prudente keuze van kenmerken kan dit probleem voorkomen. Dit is de verantwoordelijkheid van de onderzoeker of gebruiker van het algoritme en geen kenmerk van het algoritme zelf. Hoewel kenmerken van bepaalde bevolkingsgroepen input kunnen zijn voor het PMJ, heeft het PMJ niet tot doel om prognoses te maken voor specifieke bevolkingsgroepen met uitzondering van meerderjarigen en minderjarigen in verband met het verschil in het toegepaste strafrecht.

1.5 Leeswijzer

Dit rapport beschrijft op hoofdlijnen in eenvoudige termen een aantal technieken die nuttig kunnen zijn voor het PMJ. Dit rapport bevat geen technische details, maar deze zijn terug te vinden in Ter Braak et al. (2024). De lezer die niet is geïnteresseerd in de details van de verschillende technieken, kan hoofdstukken 3 t/m 6 overslaan. Wel worden een aantal gebruikte statistische begrippen, zoals gemiddelde en variantie, bekend verondersteld. De overige begrippen worden in de tekst uitgelegd (met name paragraaf 4.1). De meeste technische termen worden niet vanuit het Engels naar het Nederlands vertaald.

Om te beginnen wordt in hoofdstuk 2 een beschrijving van het huidige PMJ gegeven. Hoofdstuk 3 gaat in op alternatieve methoden om het huidige PMJ-model te schatten, terwijl hoofdstuk 4 in gaat op algoritmes die een andere specificatie van het PMJ vereisen. In hoofdstuk 5 worden methoden beschreven die betrekking hebben op een betere benutting van de dataset waarop de modellen worden geschat en getest. Hoofdstuk 6 gaat in op het combineren van methoden of steekproeven. Tot slot eindigt hoofdstuk 7 met een nabeschouwing en aanbevelingen.

2 Huidige Prognosemodel Justitiële Ketens

Het PMJ maakt prognoses voor een groot aantal justitiële organisaties in de strafrechtsketen en op het terrein van de civiele en bestuursrechtspraak. Deze organisaties zijn van elkaar afhankelijk voor hun in- en uitstroom. Deze onderlinge afhankelijkheid is een zeer belangrijk kenmerk van het PMJ. Het huidige PMJ is een combinatie van structurele modellen, voorraad-stroommodellen en tijdreeksmodellen en bestaat ongeveer uit 6.600 vergelijkingen. De coëfficiënten van de vergelijkingen worden geschat met behulp van regressieanalyse op geaggregeerde jaargegevens. Het resultaat is een empirisch model waarmee ramingen kunnen worden gegenereerd, die dienen als onderbouwing voor een deel van de begroting van het ministerie van Justitie en Veiligheid. Grofweg bestaat het model uit zeven soorten vergelijkingen:

1 instroom;
2 uitstroom;
3 doorstroom;
4 eindvoorraad;
5 beginvoorraad;
6 subcategorieën;
7 sanctieduur.

De volgende paragrafen gaan hier nader op in.

2.1 Instroom in de keten

De basis van het PMJ wordt gevormd door ontwikkelingen in de samenleving die grotendeels buiten de invloedssfeer van justitie liggen. Het uitgangspunt van het model is dat deze maatschappelijke ontwikkelingen invloed hebben op de ontwikkeling van de criminaliteit en het ontstaan en de beslechting van geschillen en daarmee dus op het beroep op de justitiële ketens.

De rol die maatschappelijke problemen volgens criminologische theorieën spelen in het ontstaan van criminaliteit en de vertaling ervan naar kwantificeerbare achtergrondfactoren wordt uitgebreid beschreven in Moolenaar et al. (2004). Het uitgangspunt is dat maatschappelijke problematiek invloed heeft op de ontwikkeling van de criminaliteit en daarmee op het strafrechtelijke gedeelte van de justitiële keten. Voorbeelden van maatschappelijke problematiek zijn:

- groepsvorming van delinquenten;
- maatschappelijke ongelijkheid;
- botsing van culturen;
- sociale instabiliteit;
- opvoeding en sociaal milieu;
- gelegenheid tot deviant gedrag;
- sociale uitstoting;
- opportunity costs (mogelijkheid, rendement en aantrekkelijkheid van andere bronnen van inkomsten in vergelijking met criminaliteit).

De economische benadering benadrukt het belang van de mogelijke opbrengsten uit criminaliteit (Becker, 1968). Criminologische theorieën proberen te verklaren welke sociale omstandigheden criminaliteit in de hand werken en welke omstandigheden ervoor zorgen dat mensen zich houden aan sociale normen en zich onthouden van crimineel gedrag (Lilly et al., 1995). Verschillen in leeftijd en geslacht vallen vaak samen met verschillen in criminaliteitsniveau, zowel aan de daderkant, maar ook aan de slachtofferkant: jonge mannen plegen de meeste delicten en jongeren zijn ook relatief vaak slachtoffers. Mogelijke oorzaken zijn fysieke en psychologische factoren, impulsiviteit en verkenning van de grenzen. Ook losse sociale verbanden en afwezigheid van sociale controle staan bekend als criminogene factoren. Macro-indicatoren in dit verband zijn een lage sociaal-economische status, woonmobiliteit, ontwrichting van gezinnen en verstedelijking. Sociaal-economische ongelijkheid kan ook een criminogene factor zijn. Macro-indicatoren hiervan zijn werkloosheid, inkomensniveau en inkomensverdeling.

Leertouwer et al. (2005, 2007) beschrijven uitgebreid hoe maatschappelijke fenomenen volgens theorieën over het ontstaan en de beslechting van geschillen samenhangen met de instroom van zaken bij de rechter en de afgifte van toevoegingen in het kader van gesubsidieerde rechtsbijstand. De maatschappelijke fenomenen die volgens de theorie een rol spelen, betreffen:

- de mate van sociale cohesie binnen de samenleving;
- probleemfrequentie, namelijk de kans op en het aantal onbevredigende transacties (te denken valt hier aan economische activiteit, arbeidsparticipatie enz.);
- financiële drempels die een rol spelen bij de inschakeling van rechtshulp;
- de organisatie van rechtshulp, die een rol speelt in de beschikbaarheid ervan.

Voor de maatschappelijke problematiek en fenomenen zijn kwantificeerbare indicatoren gezocht, die *achtergrondfactoren* worden genoemd. Deze achtergrondfactoren vormen tezamen een indicatie van bovengenoemde problemen en fenomenen. Er is geen één-op-één-relatie tussen de achtergrondfactoren en deze problemen en fenomenen. Een achtergrondfactor kan een indicator zijn voor meerdere problemen en/of fenomenen. Aan de andere kant kan een probleem of fenomeen door meerdere achtergrondfactoren worden gekarakteriseerd. De achtergrondfactoren kunnen grofweg in vier categorieën worden ingedeeld:

- demografische ontwikkelingen;
- economische ontwikkelingen;
- maatschappelijke ontwikkelingen;
- overige ontwikkelingen.

Voorbeelden van achtergrondfactoren zijn de bevolking in verschillende leeftijdscategorieën, werkloosheid, bruto nationaal product, drugsgebruik, griffierechten enz. De keuze van de achtergrondfactoren wordt naast bovengenoemde theorieën mede bepaald door beschikbaarheid en kwaliteit van gegevens.

Deze achtergrondfactoren bepalen vervolgens de instroom van zaken/dossiers/personen bij een organisatie aan het begin van de keten (bijvoorbeeld het aantal geregistreerde misdrijven door de politie). Daarbij worden verschillende typen zaken onderscheiden op basis van beleidsrelevantie, omvang en wijze waarop deze zaken door de justitiële keten stromen. In de strafrechtsketen wordt een onderscheid gemaakt tussen verschillende typen delicten en tussen volwassenen en

jongeren. In de civiele en bestuursrechtspraak worden verschillende typen conflicten onderscheiden. Per delict- of zaakstype wordt bekeken welke achtergrondfactoren mogelijk relevant kunnen zijn. Vervolgens wordt op statistische gronden beslist welke achtergrondfactoren uiteindelijk in het model voor het desbetreffende delict- of zaakstype worden opgenomen. Omdat er geen één-op-één-relatie is tussen de achtergrondfactoren en de genoemde problematiek zijn de gevonden verbanden in het model niet noodzakelijkerwijs causaal.

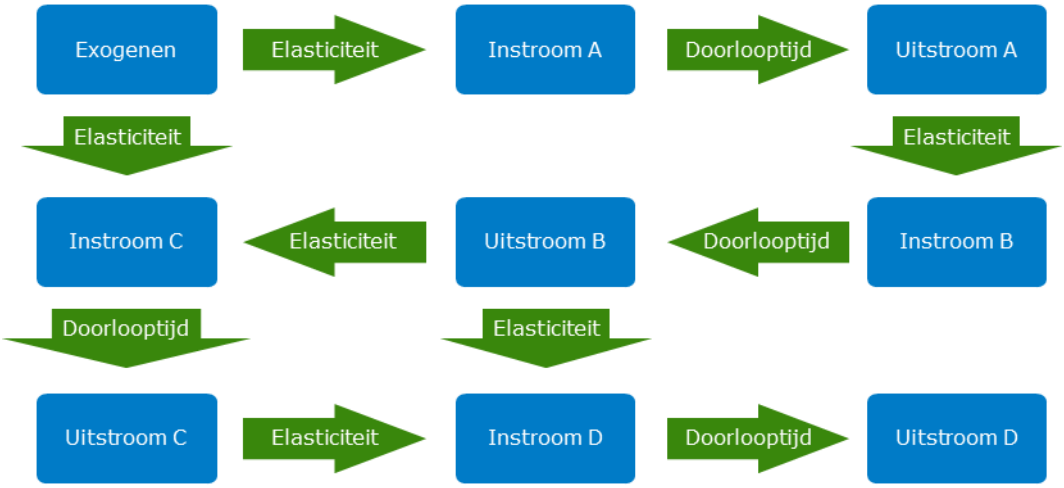
Om een prognose te maken is met name het groeitempo belangrijk, ofwel de korte termijn relatie tussen de te voorspellen instroom en de achtergrondfactoren. In sommige gevallen wordt ook de lange termijn relatie tussen te voorspellen instroom en de achtergrondfactoren verondersteld. Het groeitempo van de instroom van zaken/dossiers/personen bij een organisatie is uit te drukken in termen van elasticiteiten. Een elasticiteit geeft de procentuele verandering van een variabele aan als reactie op een verandering in een andere. Bijvoorbeeld, als de elasticiteit tussen werkloosheid en geregistreerde misdrijven geschat wordt op 0,5, dan betekent dit dat een procentuele toename van 1% van het aantal werklozen zal leiden tot een toename van 0,5% van het aantal geregistreerde misdrijven, als alle andere verklarende variabelen ongewijzigd blijven. De specificatie in elasticiteiten maakt het mogelijk om relatief eenvoudige wijze een gevoeligheidsanalyse uit te voeren. Bovendien is het schatten van de groeisnelheid in termen van een elasticiteit gelijk aan het schatten van een Cobb-Douglas-productiefunctie, die in de economische wetenschappen veel gebruikt wordt om de relatie tussen inputs en outputs van een organisatie te schatten.

2.2 Uitstroom uit en doorstroom door de keten

De uitstroom van zaken/dossiers/personen bij een organisatie, ongeacht haar plaats binnen de keten, hangt af van de beginvoorraad van het jaar, de instroom en de doorlooptijd (gemeten als een deel van een jaar). Voor sommige onderdelen is de (gemiddelde) doorlooptijd bekend, voor andere onderdelen wordt dit geschat. Bijvoorbeeld als de gemiddelde doorlooptijd van een zaak bij het Openbaar Ministerie drie maanden is, dan zal grofweg drie kwart van de zaken die dit jaar instroomt ook nog dit jaar worden afgedaan, terwijl een kwart van de zaken (met name de zaken die in het laatste kwartaal instromen) pas in het daaropvolgende jaar worden afgedaan.

Omdat met name de strafrechtsketen een aaneenschakeling is van diverse justitiële organisaties wordt in het PMJ een relatie gelegd tussen de instroom van zaken/dossiers/personen bij een organisatie ergens in het midden van de keten en de uitstroom van zaken/dossiers/personen bij de voorafgaande organisatie in de keten (zie figuur 2.1). Deze relatie is wederom geformuleerd in termen van elasticiteiten. Zo wordt bijvoorbeeld een elasticiteit geschat tussen het aantal binnenkomende strafzaken bij de Raad voor de rechtspraak en het aantal dagvaardingen door het Openbaar Ministerie. Hetzelfde principe wordt toegepast op civiele rechtspraak en bestuursrechtspraak, maar omdat deze ketens vrij kort zijn, is het aantal ketenpartners beperkt.

Figuur 2.1 Ketenconsistentie



2.3 Voorraden, subcategorieën en duur

De eindvoorraad van zaken/dossiers/personen bij een organisatie in de huidige periode is gelijk aan de beginvoorraad van de huidige periode plus de instroom minus de uitstroom. Uiteraard is de beginvoorraad van zaken/dossiers/personen in de volgende periode gelijk aan de eindvoorraad van de huidige periode. Beide vergelijkingen zijn definitievergelijkingen.

De hoofdstromen in het PMJ worden op een relatief hoog aggregatieniveau geschat. Maar soms is er behoefte aan wat meer detail. Bijvoorbeeld bij een transactie of strafbeschikking is het belangrijk om te weten of het om een geldsom, taakstraf of iets anders gaat, niet omdat het Openbaar Ministerie dit op verschillende wijzen afhandelt, maar omdat dit de routering door de rest van de strafrechtketen bepaalt. De tenuitvoerlegging van taakstraffen volgt een andere route dan de tenuitvoerlegging van financiële sancties. In dat geval worden taakstraffen en geldsommen geschat als een percentage van het totaal aantal transacties en/of strafbeschikkingen.

Bij een aantal type sancties speelt ook de duur een rol in de bepaling van de behoefte, bv. vrijheidsbenemende sancties en taakstraffen. Omdat het erg moeilijk is om te bepalen wat de factoren zijn die de duur beïnvloeden, wordt de duur per type sanctie doorgaans constant gehouden, tenzij er zwaarwegende redenen zijn om dat niet te doen (bv. een wetswijziging).

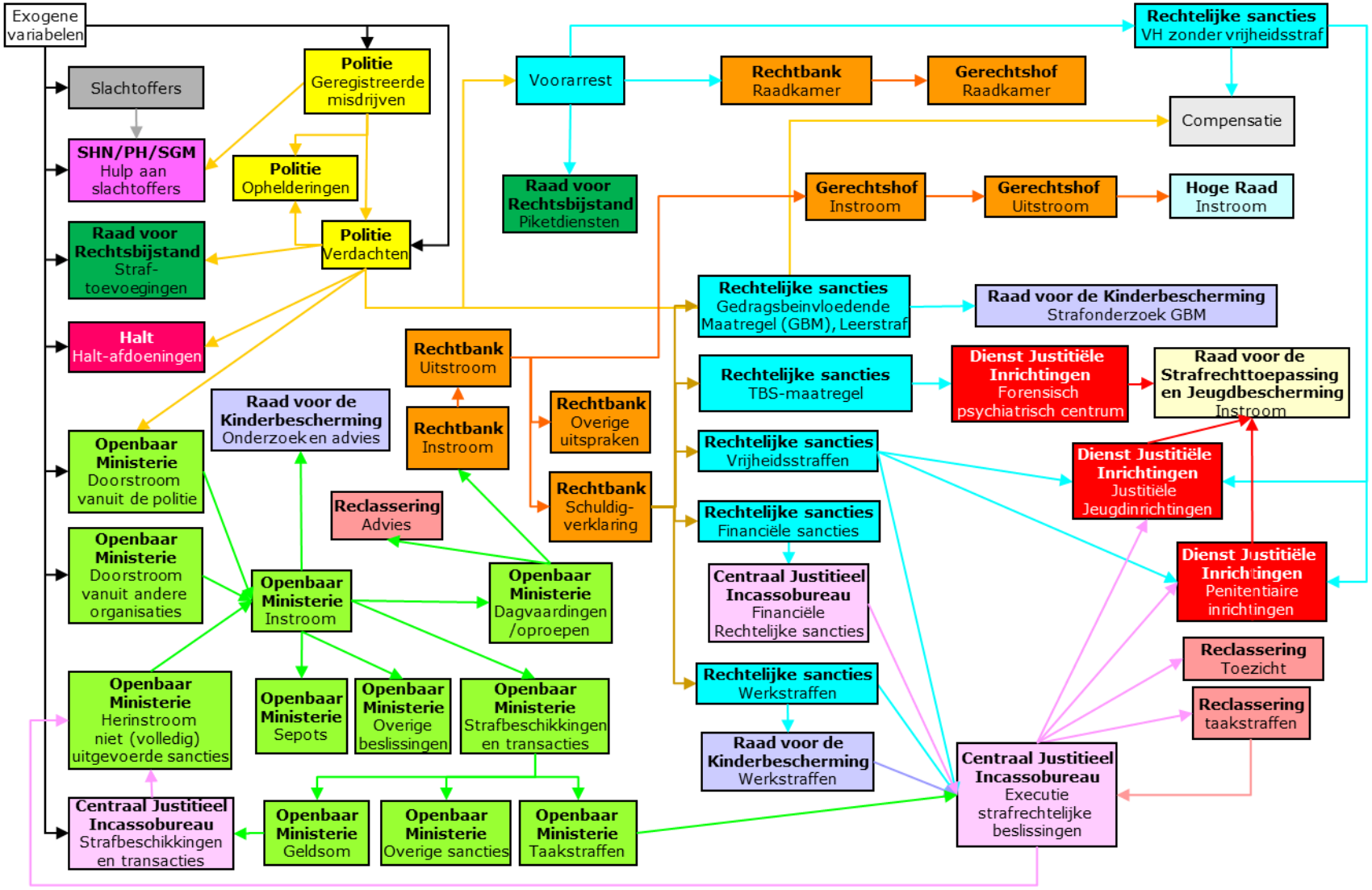
2.4 Keteneffecten

De meeste organisaties hebben verschillende soorten stromen (bijvoorbeeld rechtbankzaken en kantonzaken bij het OM en de rechtspraak; boetes, taakstraffen en coördinatie van vrijheidsstraffen bij het CJIB; gevangeniswezen, justitiële jeugdinrichtingen en vreemdelingenbewaring bij DJI, enz.). De zeven genoemde typen vergelijkingen worden zo veel mogelijk op elk type stroom bij elke organisatie voor elk delict- of zaakstype en beide leeftijdscategorieën toegepast, afhankelijk van beschikbaarheid van data en relevantie. Hierdoor ontstaat een ketenmodel. Figuren

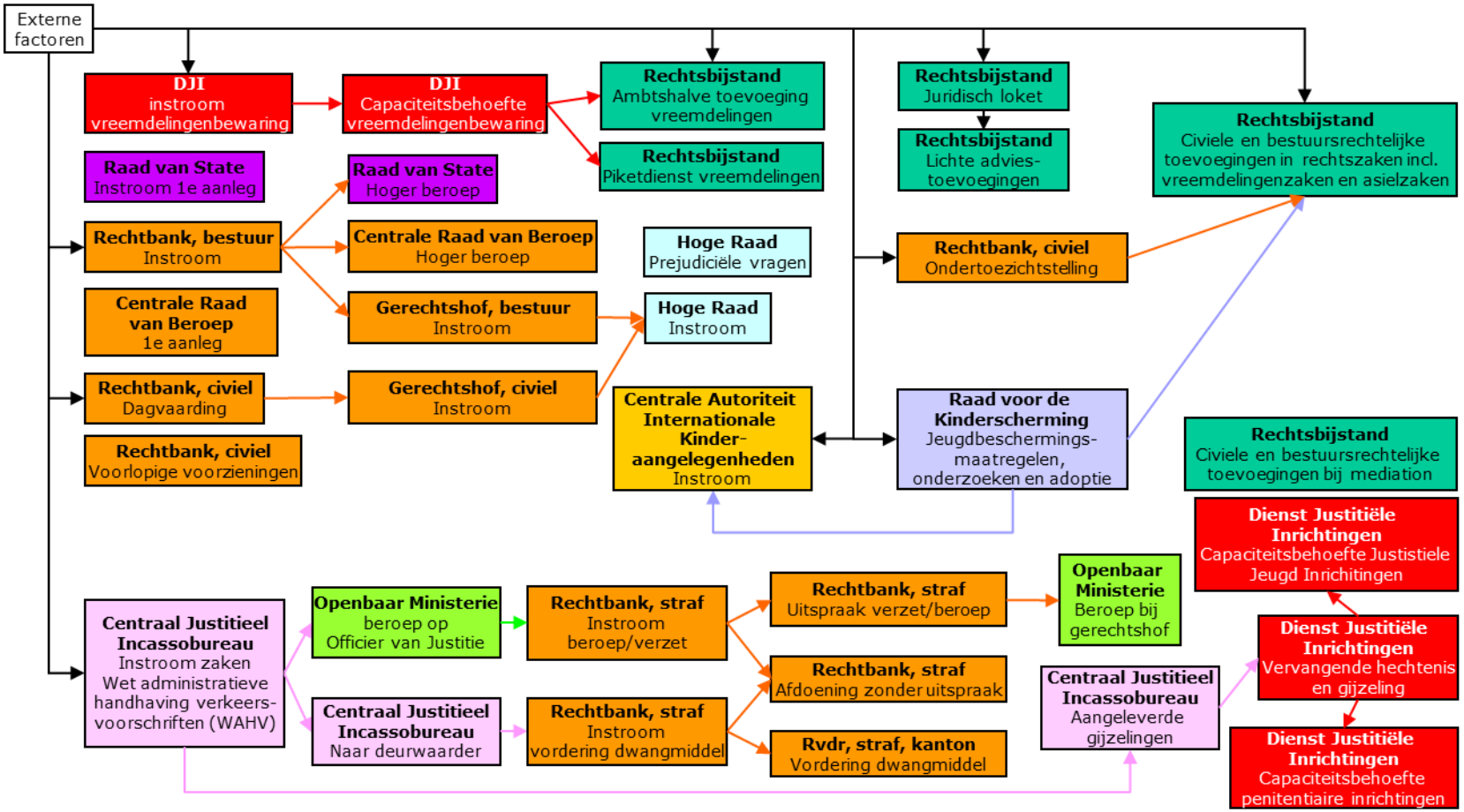
2.2 en 2.3 geven respectievelijk een schematische weergave van het strafrechtelijk deel van het model en het civiel- en bestuursrechtelijk deel van het model.

Met behulp van prognoses uit externe bronnen voor demografische, maatschappelijke en economische ontwikkelingen maakt het model een prognose voor geregistreerde criminaliteit (linksboven in figuur 2.2). Aan de hand van de prognoses voor geregistreerde criminaliteit maakt het model een prognose voor het aantal verdachten, waarmee weer een prognose wordt gemaakt voor het aantal zaken dat door het OM wordt afgehandeld. Het aantal dagvaardingen door het OM vormt de instroom van de rechterlijke macht (zittende magistratuur, ZM). De sanctie die door het OM en ZM worden opgelegd zijn de instroom bij het CJIB en DJI (rechtsonder in figuur 2.2). Op deze manier kunnen prognoses worden gemaakt voor de gehele strafrechtketen. Met behulp van prognoses uit externe bronnen voor demografische, maatschappelijke en economische ontwikkelingen maakt het model ook een prognose voor de instroom van civiele en bestuursrechtelijke zaken in eerste aanleg (zie linksboven in figuur 2.3). Deze zaken kunnen eventueel doorstromen naar het gerechtshof en de hoge raad. Ook staat onderaan in figuur 2.3 de keten voor lichte verkeerszaken beschreven. Voor meer details over het model zie Moolenaar et al. (2004) en Leertouwer et al. (2005, 2007).

Figuur 2.2 De strafrechtsketen in het PMJ



Figuur 2.3 De civielrechtelijke en bestuursrechtelijke keten in het PMJ



2.5 Kwaliteit van het model

Met elk prognosemodel is de ultieme vraag, hoe goed het model werkt. Hoewel de vraag simpel lijkt, is het antwoord niet zo eenvoudig. Uiteindelijk is een model een vereenvoudigde weergave van de werkelijkheid. De hele wereld modelleren maakt het model nodeloos complex, maar de relevante factoren die de capaciteitsbehoefte van diverse justitieonderdelen beïnvloeden moeten wel worden meegenomen. Er moet dus worden gezocht naar een balans tussen complexiteit en relevantie. Daarnaast zullen er altijd onverwachte factoren zijn die de uitkomst beïnvloeden, zoals bijvoorbeeld COVID-19. Dat maakt het vergelijken van de ramingen met de werkelijke cijfers lastig, want niet alle factoren die invloed hebben op de werkelijke cijfers, in het model zijn meegenomen. Voor de beoordeling van de kwaliteit van het model is het dus niet voldoende om alleen maar naar de cijfers te kijken. In paragraaf 2.5.1 wordt kort beschreven wat de voor- en nadelen van PMJ zijn. Paragraaf 2.5.2 gaat in op eerdere (externe) evaluaties. In paragraaf 2.5.3 wordt naar de cijfers gekeken.

2.5.1 Voor- en nadelen

Het grote voordeel van het PMJ is de ketenconsistentie van de ramingen. Voor 1998 werden de ramingen voor de diverse justitieonderdelen los van elkaar gemaakt, waardoor de ramingen niet onderling consistent waren. Hierdoor waren bottlenecks in de keten ook minder goed zichtbaar. Met de komst van PMJ is de onderlinge afhankelijkheid van de justitiële organisaties beter in kaart gebracht.

Het PMJ wordt vaak als complex ervaren, maar deze complexiteit komt deels voort uit de complexiteit van de strafrechtsketen zelf. Zo zijn er in de strafrechtsketen heel veel mogelijkheden tot herinstroom, bijvoorbeeld als gevolg van het niet (volledig) betalen van een boete, het niet (volledig) uitvoeren van een taakstraf of vrijheidsstraf, het niet voldoen aan de voorwaarden van een voorwaardelijk sepot of straf, enz. Hierdoor stromen zaken die uitgestroomd lijken, toch weer opnieuw in in de keten. Het PMJ weerspiegelt slechts deze complexiteit.

Vanuit een statistisch/wiskundig oogpunt zijn de vergelijkingen in het PMJ relatief eenvoudig. Per module is goed uitlegbaar en achterhaalbaar wat er precies gebeurt. Maar door de stapeling van modules (omdat het nu eenmaal een ketenmodel is) zijn de ramingen achter in de keten soms lastiger uit te leggen, wat vaak ook als complex wordt ervaren. Bijvoorbeeld, de ramingen van de capaciteitsbehoefte van DJI zijn afhankelijk van de ramingen voor de rechtspraak, die op hun beurt weer afhankelijk zijn van de ramingen voor het OM, die op hun beurt weer afhankelijk zijn van de ramingen voor de politie. Dit betekent dat in het PMJ geen directe relatie is tussen de capaciteitsbehoefte bij DJI en de criminaliteit, maar wel indirect. Het is mogelijk om een directe relatie te leggen, maar daarmee gaat wel de ketenconsistentie verloren. Omdat PMJ vanuit statistisch/wiskundig oogpunt relatief eenvoudig is, is het model makkelijk te implementeren met standaard software. Omdat het PMJ zeven jaar vooruit voorspelt (gelijk aan de begrotingshorizon) kan niet worden volstaan met uitsluitend data-analyse. Er moeten theoretische aannames worden gemaakt over de onderlinge relaties tussen justitieonderdelen voor de komende zeven jaar. Hiervoor is ook domeinkennis nodig. Omdat alleen gebruik wordt gemaakt van geaggregeerde data zijn er geen privacy issues. In de ramingen worden geen specifieke bevolkingsgroepen onderscheiden (m.u.v. jeugd en volwassenen i.v.m. jeugdstrafrecht), zodat het PMJ ook niet discrimineert. Omdat het PMJ een ketenmodel

is, kan het ook worden gebruikt voor simulaties. Het PMJ kan bijvoorbeeld uitrekenen wat het effect is achteraan in de strafrechtketen van het verschuiven van prioriteiten bij het OM en/of de politie.

2.5.2 Evaluaties

Het PMJ is voortgekomen uit het Jukeboxmodel.[1] Dit is in 1998 ontwikkeld door het Sociaal en Cultureel Planbureau (SCP) en daarna overgedragen aan het toenmalige Wetenschappelijk Onderzoek- en Documentatiecentrum. In 2003 is begonnen met de uitbreiding van het Jukeboxmodel richting het huidige PMJ. De modellen zijn inmiddels dertien keer geëvalueerd:

- SEO: Theeuwes & De Winter (1998);
- KPMG/BEA (1998);
- IVA: Spapens et al. (2001);
- DSRS (2002);
- Nyfer: Bomhoff et al. (2002);
- SEO: Biermans & Van Leeuwen (2003);
- APE: Goudriaan (2004);
- SEO: Felsö et al. (2006);
- Regioplan: Bont et al. (2009);
- WODC: Moolenaar et al. (2009);
- APE: Everhardt et al. (2016);
- WODC: Moolenaar et al. (2018);
- Nobis Policy Lab: De Poot et al. (2020).

De algemene conclusie van de evaluaties is dat de betrouwbaarheid van het model niet perfect is, maar wel het hoogst haalbare en dat het model complex is. Dat laatste is onvermijdbaar aangezien met name de strafrechtsketen zelf zeer complex is. De conclusie van de op één na laatste externe evaluatie van Everhardt et al. (2016) luidde dat

> "(…) er geen aanwijzingen zijn dat het model fundamenteel zou moeten worden herzien vanwege nieuwe econometrische en/of criminologische inzichten. (…) Het model zit econometrisch goed in elkaar."

Wel adviseerden zij om de analyse van de voorspelfouten uit 2009 te herhalen en de mogelijkheid van hoogfrequente data te onderzoeken. Het resultaat hiervan is te lezen in Moolenaar et al. (2018):

> "Een foutieve uitgangswaarde kan sterk doorwerken in de ramingen. Het blijkt dat het laatste bekende realisatiejaar achteraf nog gemiddeld met zo'n 3% naar beneden wordt bijgesteld. (…) De inschatting van beleidseffecten van nieuw beleid en/of nieuwe wetgeving door de beleidsdirecties en uitvoeringsorganisaties heeft doorgaans een opwaarts effect op de voorspelfout. (…) De voorspelfouten in de beleidsrijke PMJ-ramingen zijn tot en met het eerste begrotingsjaar (drie jaar vooruit) kleiner of vrijwel gelijk aan die van alternatieve eenvoudige tijd-reeksmodellen, zoals constant houden en trendextrapolatie. (…) Op de lange termijn blijkt het PMJ relatief minder goed te voldoen. Constant houden op laatste realisatiejaar of eerste of tweede begrotingsjaar is dan altijd beter. Maar

[1] Jukebox staat voor Justitieketenmodel box 1, waarbij box 1 het volwassenstrafrecht betrof. Box 2 had betrekking op het jeugdstrafrecht.

> tijdreeksmodellen doen het nog slechter dan PMJ. (..) Het PMJ heeft bovendien als voordeel dat het de samenhang bevordert tussen de ramingen van diverse onderdelen van de justitiële keten, en dat het simulaties van bepaalde beleidseffecten en van snel wijzigende economische omstandigheden mogelijk maakt. (…) Eenvoudige recepten voor verbetering op dit gebied zijn er niet. (..) Voor de nabije toekomst lijkt de bruikbaarheid van big data (microdata uit administratieve bestanden, van sociale media, dark web en/of internet-of-things) ten behoeve van het PMJ om uiteenlopende redenen beperkt. Wel is het mogelijk om sommige technieken die vaak worden toegepast op big data, toe te passen op de data die momenteel wel beschikbaar zijn. Dit biedt echter geen garantie op betere ramingen."

De Poot et al. (2020) hebben de conclusie van Everhardt et al. (2016) overgenomen en hebben zich derhalve meer op het proces rondom het PMJ gericht. Zij hebben de wensen van de eindgebruikers van het PMJ in kaart gebracht. De wensen kunnen grofweg in negen categorieën worden ingedeeld (tussen haakjes de gemiddelde score op een schaal van 1 tot en met 5):

1 proceswensen (1.8);
2 de wens voor frequentere herijkingen (updates) t.a.v. PMJ (3.0);
3 analysewensen t.a.v. mogelijke causale verbanden (2.8);
1 wensen t.a.v. dilemma's in capaciteitsprognoses (2.4);
2 gegevens voor prognoses combineren (2.0);
3 wensen voor het tussentijds wijzigen van de prijs (1.5);
4 wensen t.a.v. uitbreiden van informatie in PMJ (1.2);
5 het opnemen van vermeden ketenbelasting als additionele grootheid in het PMJ (1.0);
6 onderzoekswensen t.a.v. het PMJ (2.0).

Uit het rapport van De Poot et al. (2020) komt niet naar voren dat gebruikers behoefte hebben aan een radicaal ander model. De voorbeelden die door enkelen genoemd werden bij punt 1, 2 en 5 (boete- en transactiemodel, MPP, asielramingen) lijken relatief eenvoudige (tijdreeks)modellen vergeleken met het PMJ. Maar de gebruikte technieken in deze modellen verschillen niet wezenlijk van de gebruikte technieken in het PMJ (namelijk regressieanalyse, tijdreeksanalyse, stroom-voorraadmodel, alles op geaggregeerd niveau). De complexiteit van het PMJ-model zit 'm dan ook niet zo zozeer in de gebruikte technieken, maar komt vooral voort uit de complexiteit van de strafrechtsketen zelf.

2.5.3 Voorspelkwaliteit

Het WODC en de Rvdr monitoren de kwaliteit van de ramingen regelmatig, via bepaling achteraf van de voorspelfouten in de ramingen. Dit houdt in dat ramingen uit het verleden vergeleken worden met inmiddels bekende gerealiseerde waarden. De laatste analyse van de voorspelfouten is uitgevoerd in Moolenaar et al. (2021). In principe zou eenzelfde analyse uitgevoerd kunnen worden over de meest recente periode, maar door de COVID-19-pandemie is gedurende twee à drie jaar een uitzonderlijke situatie ontstaan en dus zegt een vergelijking van de ramingen met realisaties voor de jaren 2020 en 2021 niets over de kwaliteit van het model. Daarnaast is een complicerende factor dat het PMJ-model beleidsneutrale ramingen berekent, terwijl de realisaties beleidseffecten bevatten en dus per definitie beleidsrijk zijn. Daarom is in Moolenaar et al. (2021) alleen de voorspelfout van de beleidsrijke

ramingen gepresenteerd. Verder blijkt dat de (voorlopige) realisatiecijfers waarmee de prognoses berekend worden, achteraf nog veelvuldig worden bijgesteld (zie tabel 8.2 in Moolenaar et al., 2021). Het laatste bekende realisatiejaar wordt achteraf nog gemiddeld met zo'n 8% bijgesteld. Vermoedelijk worden eerdere jaren ook bijgesteld maar dit is niet onderzocht. De gemiddelde bijstelling van de voorlopige realisatie-cijfers over het lopende jaar (het jaar waarin de ramingen worden berekend) is nog ongeveer even groot, afgerond ook 8%. Kortom, per saldo wordt doorgaans van hogere aantallen uitgegaan dan achteraf het geval is. Deze fout werkt door in de hele prognose.

Uit een eerder analyses van Moolenaar et al. (2009 en 2018) bleek dat er gemiddeld sprake is van enige overschatting, maar niet systematisch. De voorspelfouten in de beleidsrijke PMJ-ramingen zijn tot en met het eerste begrotingsjaar (drie jaar vooruit) kleiner of vrijwel gelijk aan die van alternatieve eenvoudige tijdreeksmodellen. Op de korte termijn levert de investering in het PMJ dus winst op. Op de lange termijn blijkt het PMJ relatief minder goed te voldoen dan constant houden. Maar tijdreeksmodellen doen het nog slechter dan PMJ.

2.6 PMJ in de toekomst

Uit eerdere externe evaluaties blijkt dus dat het PMJ goed in elkaar zit (Everhardt et al., 2016) en dat gebruikers geen behoefte hebben aan een radicaal ander model (De Poot et al., 2020). Maar het huidige PMJ is ontworpen in een periode waarin de beschikbaarheid van microdata beperkt was en een aantal technieken vaak wel theoretisch bekend waren, maar niet konden worden geïmplementeerd vanwege beperkingen in computertechnologie. Daarom is het zinvol om te onderzoeken of in de afgelopen jaren ontwikkelingen zijn geweest op het gebied van data en prognosetechnieken, die meer of andere inzichten kunnen bieden.

In het vervolg van dit rapport zal worden bekeken welke alternatieve methoden er zijn voor het maken van prognoses, wat voor data daarvoor nodig is en in hoeverre ze zouden kunnen worden toegepast zonder de voordelen van het huidige PMJ teniet te doen, in het bijzonder de ketenconsistentie. Deze methoden vallen grofweg uiteen in vier categorieën:

- methoden die de parameters van het huidige PMJ-model op een andere manier schatten (zie hoofdstuk 3);
- methoden waarbij (delen van) het model anders worden gespecificeerd (zie hoofdstuk 4);
- methoden die betrekking hebben op een betere benutting van de dataset waarop de modellen worden geschat en getest (zie hoofdstuk 5);
- combinatie van methoden en/of steekproeven (zie hoofdstuk 6).

Voor de eenvoud wordt in deze hoofdstukken ervanuit gegaan dat er slechts één variabele wordt voorspeld. De generalisatie naar het voorspellen van meerdere variabelen tegelijkertijd wordt niet in dit rapport behandeld.

Vooralsnog is deze verkenning een zuiver theoretische aangelegenheid. In de aanbevelingen van dit rapport zal worden aangegeven wat de meest belovende technieken zijn, die in aanmerking komen voor een pilot in een vervolgonderzoek. Dit betekent dat er in dit rapport geen uitspraken kunnen worden gedaan over de

voorspelkwaliteit van de alternatieve technieken. Er is dus geen garantie dat de alternatieve technieken ook daadwerkelijk betere ramingen opleveren dan het PMJ. De toekomstige pilots zullen hierover meer uitsluitsel kunnen geven.

3 Aanscherping van het huidige PMJ

Dit hoofdstuk richt zich op diverse alternatieve methoden om de parameters van het huidige PMJ-model te schatten. Het PMJ bevat endogene en exogene variabelen. Een exogene variabele is een variabele waarvan de waarde buiten het model wordt bepaald en aan het model wordt opgelegd. Afhankelijk van de wetenschappelijke discipline worden exogene variabelen ook wel 'features', kenmerken, covariaten, inputs, onafhankelijke variabelen of verklarende variabelen genoemd. Een endogene variabele daarentegen is een variabele waarvan de (nog niet geobserveerde) waarde wordt geschat door het model. Een endogene variabele wordt ook wel een label, output, 'target', doel(variabele) of afhankelijke variabele genoemd.

Omdat met geaggregeerde data wordt gewerkt zijn de meeste endogene variabelen in het PMJ continu en bestaat het PMJ voornamelijk uit lineaire regressies. Lineair wil zeggen dat het model lineair is in de parameters maar niet noodzakelijkerwijs in de variabelen. Dit betekent dat zowel de endogene als de exogene variabelen een niet-lineaire transformatie kunnen hebben ondergaan, zolang de relatie ertussen maar wel lineair is (dat wil zeggen een optelsom). Daarom komen in dit hoofdstuk alleen verschillende vormen van lineaire regressie aan bod. Het overzicht in dit hoofdstuk is geenszins uitputtend. Er zijn vele regressie-algoritmes maar in dit hoofdstuk is geselecteerd op algoritmes die mogelijk relevant kunnen zijn voor het PMJ-model omdat ze een oplossing bieden voor veel voorkomende problemen.

3.1 Lineaire regressie

Het doel van een lineaire regressie is om een lijn te vinden die zo goed mogelijk de relatie tussen de endogene variabele en exogene variabele(n) beschrijft en daarmee de waarde van de endogene variabele kan voorspellen op basis van de waarden van de exogene variabele(n). Om de parameters van het model te schatten kunnen verschillende algoritmes worden toegepast. De meest eenvoudige methode is de lineaire of gewone kleinste kwadraten methode (zie paragraaf 3.1.1).
Als er uitschieters in de data zijn (d.w.z. grote residuen) of als er sprake is van overfitten (d.w.z. dat het belang van één of meer exogene variabelen wordt overschat) of de dataset heel klein is of scheef verdeeld of als een deel van de gegevens waarin men geïnteresseerd is niet wordt waargenomen of als de storingstermen niet normaal verdeeld zijn, kan de gewone kleinste kwadratenmethode resulteren in een empirisch model dat geen goede representatie is van de onderliggende statistische relatie. Deze problemen kunnen worden oplost door een ander algoritme te kiezen (zie paragraaf 3.1.2 tot en met 3.1.7).

3.1.1 Gewone kleinste kwadraten

De gewone kleinste kwadraten methode ('ordinary least squares', OLS) is het meest eenvoudige algoritme om de parameters van een theoretisch model te schatten. Om deze reden wordt het ook veel toegepast, ook in het huidige PMJ-model. Bij de kleinste kwadratenmethode wordt de som van de kwadraten van het verschil tussen de werkelijke waarde en de geschatte waarde van de endogene variabele geminimaliseerd. Alle exogene variabelen, parameters en waarnemingen worden gelijk behandeld. In de meest simpele vorm wordt een endogene variabele geregresseerd op

een constante. Met de kleinste kwadratenmethode is dit equivalent aan het berekenen van het gemiddelde van de waarden van een endogene variabele.

Het grote voordeel van OLS is dat het breed toepasbaar is en relatief eenvoudig te implementeren. Omdat de methode een analytische oplossing heeft, is de benodigde rekentijd gering. Het kan worden toegepast op zowel micro-data als geaggregeerde data. De methode kan zowel voor korte als (middel)lange termijn prognoses worden gebruikt. Technisch gezien is de methode transparant en zijn de uitkomsten goed herleidbaar naar de exogene variabelen. Het is ook inhoudelijk goed uitlegbaar zolang de data niet getransformeerd zijn en als de tekens van de parameters conform verwachtingen zijn. De methode is nog steeds inhoudelijk uitlegbaar als de data dusdanig zijn getransformeerd dat de transformatie een linearisatie van een niet-lineaire relatie representeert. Zo bestaat het huidige PMJ grotendeels uit lineaire regressies op getransformeerde data, waarbij de resulterende vergelijkingen een linearisatie van een Cobb-Douglas productiefunctie voorstellen en de parameters elasticiteiten zijn.

Wel moeten de data aan een aantal voorwaarden voldoen, zoals homoscedasticiteit (variantie constant), geen seriële correlatie, geen uitschieters, geen systematische meetfouten in de exogene variabelen, ongecorreleerde exogene variabelen, geen afgeknotte of gecensureerde variabelen en de storingsterm moet normaal verdeeld zijn. In het geval niet aan één of meerdere voorwaarden wordt voldaan kan een schatter gekozen worden uit de hiernavolgende subparagrafen.

3.1.2 Correctie voor niet-constante variantie

Als de variantie niet constant is, is er sprake van heteroscedasticiteit. In dit geval kan de gewone kleinste kwadraten algoritme worden aangepast, zodanig dat hiermee wordt rekening gehouden. Dit heeft geen effect op de weging van de exogene variabelen, maar alleen op de schatting van de variantie en daarmee op de onzekerheidsmarges.

3.1.3 Correctie voor slecht gemeten of sterk gecorreleerde exogene variabelen

Soms zijn exogene variabelen sterk gecorreleerd met elkaar, bijvoorbeeld inkomen en opleidingsniveau. Of exogene variabelen worden slecht gemeten: respondenten zijn bijvoorbeeld niet altijd bereid vragen over inkomen te beantwoorden. In deze gevallen kan een instrumentele variabelen (IV) schatting uitkomst bieden. Bij een instrumentele variabelen schatting worden één of meerdere extra exogene variabelen die gecorreleerd zijn met de problematische exogene variabelen, maar ongecorreleerd met de storingsterm, als zogenoemde instrumenten in de lineaire regressie (maar niet in het model) meegenomen. Maar een veelvoorkomend probleem bij deze methode is om geschikte instrumenten te vinden.

3.1.4 Correctie voor overfitting

Ridge regressie (ook wel L_2 regularisatie genoemd) en de 'least absolute shrinkage and selection operator' (LASSO, ook wel L_1 regularisatie genoemd) en 'Elastic Net' (een combinatie van LASSO en ridge regressie) zijn variaties op de kleinste kwadratenmethode, waarbij een straf wordt gezet op overfitting. Hiertoe worden alle parameters verminderd met eenzelfde factor en de sterkte van deze regularisatie wordt geregeld middels een hyperparameter (zie box 3.1 voor de definitie). Als de

hyperparameter heel groot is, zullen deze methoden de parameters van het model heel klein schatten. Als de hyperparameter gelijk aan nul is, is er geen bestraffing en is het algoritme weer gelijk aan een gewone lineaire regressie. Uiteindelijk wordt de hyperparameter gekozen die tot de kleinste voorspelfout leidt.[2] Merk op dat deze bestraffing geen onderdeel van het model is, maar alleen van het algoritme om de parameters te schatten.

Box 3.1 Parameters en hyperparameters

De parameters zijn onderdeel van het model en worden door een algoritme op basis van de data geschat of geleerd. Parameters worden ook wel coëfficiënten of gewichten genoemd. Parameters moeten niet worden verward met hyperparameters. Hyperparameters controleren het algoritme en worden gebruikt om de optimale complexiteit van het empirische model te bepalen. De waarde van een hyperparameter wordt gekozen voordat een lerend algoritme wordt getraind. Voorbeelden van hyperparameters zijn het aantal verborgen lagen in een neuraal netwerk, het aantal iteraties dat een algoritme mag doorlopen, het aantal clusters, het aantal exogene variabelen in een model enz.

Bij een ridge regressie wordt de hyperparameter toegepast op som van de kwadratische coëfficiënten en bij LASSO wordt de hyperparameter toegepast op som van de absolute waarde van de coëfficiënten. Bij elastic net wordt de hyperparameter toegepast op een gewogen combinatie van kwadratische coëfficiënten en de absolute waarde van de coëfficiënten. Bij een ridge regularisatie worden coëfficiënten proportioneel verkleind, terwijl de LASSO-regularisatie de coëfficiënten van minder relevante exogene variabelen naar nul duwt, waardoor er een selectie effect ontstaat.

De voor- en nadelen van deze methoden komen grotendeels overeen met de gewone kleinste kwadratenmethode. Ridge regressie heeft een analytische oplossing, maar kost meer rekentijd omdat voor elke waarde van de hyperparameter het model opnieuw geschat moet worden. LASSO en Elastic Net hebben geen analytische oplossing. Voorwaarde is dat alle variabelen gestandaardiseerd zijn, dat wil zeggen op dezelfde schaal gemeten (zie box 3.2). Een nadeel is dat regularisatie onzuiverheid (‘bias’) introduceert. Bias is een vertekening of afwijking van het juiste resultaat die een systematische oorzaak heeft en dus niet te wijten is aan toevallige effecten of aan de steekproef.[3] Hierdoor kan een statistische toets op significantie van de parameters niet meer worden uitgerekend.

[2] Op de validatieset (zie paragraaf 4.1.2).
[3] Zie ook paragraaf 4.1.3.

Box 3.2 Schaling van de variabelen: standaardisatie versus normalisatie

Soms is het belangrijk dat variabelen op dezelfde schaal worden gemeten. Dat is bijvoorbeeld het geval voor algoritmen met afstandsmaatstaven omdat deze sterk worden beïnvloed door het verschil tussen twee variabelen of voor algoritmen waarbij de omvang van de parameters mede afhankelijk is van de meeteenheid van de variabelen. Er zijn meerdere technieken voor het schalen van variabelen, waarvan we hier twee bespreken, namelijk normalisatie en standaardisatie. Normalisatie schaalt de variabelen zodanig dat de waarden van de geschaalde variabelen lopen van nul tot en met één. Dit wordt ook wel min-max schaling genoemd. Bij standaardisatie wordt de variabele dusdanig getransformeerd zodat de getransformeerde variabele een gemiddelde van nul heeft en een variantie van één. Dit wordt ook wel de z-score of z-z-score normalisatie genoemd. Merk op dat de waarden van gestandaardiseerde gegevens niet beperkt zijn tot een bepaald bereik in tegenstelling tot genormaliseerde gegevens.
Of normalisatie dan wel standaardisatie moet worden toegepast, is afhankelijk van de verdeling van de gegevens. Over het algemeen verbetert normalisatie de prestaties van het prognosemodel wanneer de gegevens niet normaal verdeeld zijn en/of de algoritmen geen aannames doen ten aanzien van de verdeling van de gegevens. Als de gegevens normaal verdeeld zijn, kan standaardisatie de prestaties van het prognosemodel verbeteren. Maar er is geen duidelijk gedefinieerde regel om te beslissen wanneer de gegevens moeten worden genormaliseerd of gestandaardiseerd. In geval van twijfel kunnen beide methoden worden toegepast en de resultaten worden vergeleken.

3.1.5 Correctie voor uitschieters

Er zijn een aantal regressie-algoritmes ontworpen om minder gevoelig voor uitschieters te zijn. Dit wordt robuuste kleinste kwadraten genoemd en het is een vorm van gewogen kleinste kwadraten. Drie belangrijke algoritmes zijn de M-schatting (Huber, 1973), S-schatting (Rousseeuw en Yohai, 1984) en MM-schatting (Yohai 1987). M-schatting richt zich op uitschieters in de endogene variabele. De S-schatting richt zich op uitschieters in de exogene variabelen. MM-schatting is een combinatie van S-schatting en M-schatting en richt zich op de uitschieters in zowel de endogene als de exogene variabele(n). Zowel bij M-schatting als bij S-schatting wordt het *geschaalde* verschil tussen de werkelijke waarde en de geschatte waarde van de endogene variabele geminimaliseerd. De schaling is een niet-lineaire functie van dit verschil. Ook hier geldt dat de schaling geen onderdeel van het model is, maar alleen van het algoritme om de parameters te schatten.

Omdat er geen analytische oplossing is, worden bij robuuste kleinste kwadraten de parameters iteratief bepaald. Dit is een rekenintensieve procedure, met name bij S-schatting en MM-schatting. Verder gelden dezelfde voor- en nadelen van de gewone kleinste kwadratenmethode. Overigens kan bij incidentele uitschieters in de data ook worden gekozen voor een simpelere oplossing, namelijk het opnemen van een dummy variabele in de vergelijking, waarna deze met gewone kleinste kwadraten wordt geschat.

3.1.6 Niet normaal verdeelde storingsterm

Bij gewone lineaire regressie wordt aangenomen dat de storingstermen een normale verdeling volgen. Als dat niet het geval is, dan kan de maximale

aannemelijkheidsschatter (‘maximum likelihood estimator’, MLE) worden gebruikt. Hierin wordt de vorm van de verdeling van de storingstermen expliciet meegenomen en dat kan dus een andere verdeling dan een normaalverdeling zijn. Als er geen bijzonderheden in de data zijn en de storingstermen zijn normaal verdeeld, dan geven de gewone kleinste kwadraten schatter en de MLE hetzelfde resultaat.

3.1.7 Correctie voor niet-waargenomen data

Soms wordt een deel van de gegevens waarin we geïnteresseerd zijn, niet waargenomen. Als voor een deel van de gewenste dataset zowel de waarden van de endogene variabelen als van de exogene variabelen niet beschikbaar zijn, spreken we van afgeknotte data. Feitelijk is dan slechts een subset van de gewenste dataset beschikbaar. Als voor een deel de gewenste dataset wel de waarden van de exogene variabelen beschikbaar zijn, maar niet de exacte waarden van de endogene variabelen, dan spreken we van gecensureerde data. Een model dat geschat wordt op basis van gecensureerde data wordt ook wel een Tobit model genoemd. Zie box 3.3 voor een voorbeeld.

Box 3.3 Voorbeeld van afgeknotte en gecensureerde data

Stel we zijn geïnteresseerd in het aantal delicten dat door veelplegers wordt gepleegd en hebben daarvoor een steekproef van veelplegers beschikbaar. De politie definieert een veelpleger als iemand die tien of meer delicten heeft gepleegd. Het aantal delicten per veelpleger kan worden geschat op basis van de kenmerken van de dader. Maar omdat de steekproef geen waarnemingen bevat waarbij het aantal delicten minder dan tien is, is hier sprake van afgeknotte data. Stel dat de steekproef aangevuld kan worden met de kenmerken van daders die minder dan tien delicten hebben gepleegd, zonder dat bekend is hoeveel delicten dat dan precies zijn. In dat geval spreken we van gecensureerde data omdat wel de waarden van de exogene variabelen bekend zijn, maar niet alle waarden van de endogene variabele.

In een steekproef met afgeknotte of gecensureerde data zijn de storingstermen van het lineaire regressiemodel niet meer normaal verdeeld en moeten de parameters van het model met maximale aannemelijkheid worden geschat, waarbij de aangepaste verdeling expliciet kan worden meegenomen.

3.2 Bayesiaanse benadering van lineaire regressie

De bayesiaanse benadering van lineaire regressie kan handig zijn als de dataverzameling weinig gegevens bevat of als de gegevens scheef zijn verdeeld. Het doel van bayesiaanse regressie is niet alleen om een schatting van de parameters van het theoretisch model te vinden, maar ook om een kansverdeling voor deze parameters te vinden. Niet alleen de endogene variabele maar ook de parameters worden verondersteld uit een kansverdeling voort te komen, meestal de normaalverdeling. Dit in tegenstelling tot de lineaire regressie waar de endogene variabele wordt voorspeld uit de waardes van de exogene variabelen. In het geval van een oneindig aantal waarnemingen, convergeren de met bayesiaanse regressie gevonden waarden voor de parameters naar de waarden verkregen uit een lineaire regressie.

Voor veel algoritmes zijn vaak grote dataverzamelingen nodig. Maar de bayesiaanse regressie is zeer effectief als de dataverzameling klein is. De bayesiaanse benadering kan worden gebruikt met beperkte voorkennis over de dataverzameling. Dit maakt deze methode zeer geschikt voor modellen die leren op basis van 'real time'. Bij veel andere algoritmes moet de volledige dataset beschikbaar zijn voordat het trainen van het model kan beginnen. Verder kan bayesiaanse regressie omgaan met scheve verdelingen. Met bayesiaanse regressie kan ook makkelijk voorkennis in het empirisch model worden opgenomen. De invloed daarvan zal sterker zijn naarmate de gegevensverzameling kleiner is. Maar het proces van bayesiaanse regressie kan tijdrovend zijn. En als er grote hoeveelheden data beschikbaar zijn, loont de bayesiaanse benadering vaak niet en werkt de reguliere benadering efficiënter. Ook voor bayesiaanse regressie geldt dat het model lineair in de parameters moet zijn.

De keuze van bepaalde parameters in een bayesiaanse regressie kan een regulariserend effect hebben. Onder bepaalde voorwaarden komt de ridge regressie of LASSO overeen met bayesiaanse regressie. Maar een bayesiaanse regressie is niet hetzelfde als een ridge regressie of LASSO, omdat ridge regressie en LASSO varianten zijn op lineaire regressie en de bayesiaanse benadering een algemene manier is om statistische modellen te definiëren en te schatten. En dit kan op verschillende type modellen worden toegepast.

3.3 Samenvatting en implicaties voor het PMJ

Lineaire regressie is een eenvoudig en daarom veelvuldig toegepast algoritme, ook in het huidige PMJ-model. Als de data geen noemenswaardige problemen bevat, is OLS een goed algoritme. Als er wel problemen zijn, dan kan afhankelijk van het type probleem voor een aangepast algoritme worden gekozen. De eigenschappen van de diverse algoritmes zijn in tabel 3.1 nog eens op een rijtje gezet.

Het huidige PMJ-model werkt uitsluitend met geaggregeerde tijdreeksdata. Er zijn geen aanwijzingen dat uitschieters, slecht gemeten of onderling gecorreleerde exogene variabelen, afgeknotte of gecensureerde data en kleine of scheef verdeelde steekproeven een groot probleem vormen. Als twee exogene variabelen sterk gecorreleerd zijn, kan dat ook worden opgelost door slechts één van de twee variabelen in het model op te nemen. Omdat met geaggregeerde data wordt gewerkt, zijn er vaak voldoende alternatieven, dit in tegenstelling tot microdata waarbij men gebonden is aan wat er in het registratiesysteem staat of aan de vragen die in een enquête zijn gesteld. Uitschieters komen weliswaar voor, maar vaak heeft dat ook een aanwijsbare (incidentele) oorzaak. In dat geval kan dit probleem ook worden opgelost door een dummy variabele in het model op te nemen. Het voordeel daarvan is dat het geen gevolgen heeft voor de prognoses, in tegenstelling tot een algoritme waarbij de uitschieters worden gedempt. In de data die worden gebruikt voor PMJ komen gecensureerde data niet veel voor, met uitzondering van de duur van het verblijf in een justitiële inrichting. Dit is echter een bijzonder geval van gecensureerde data die in paragraaf 4.4.3 zal worden besproken.

Een algoritme dat mogelijk wel interessant is om op in te zoomen is elastic net regularisatie. Dit zou tot meer inzicht kunnen leiden in de invloed van de exogene variabelen in de verschillende submodellen. Merk op dat de exogene variabelen in een submodel, de endogene variabelen (ofwel de uitkomsten) van een submodel voor een voorafgaande ketenpartner kunnen zijn.

Tabel 3.1 Kenmerken van lineaire regressie-algoritmes

	OLS	IV	Elastic net	Robuust	Maximum likelihood	Afgeknotte regressie	Tobit	Bayes
Algoritme	regressie	regressie	regressie	regressie	regressie	regressie	regressie	regressie
Endogene	continu	continu	continu	continu	continu	continu	continu	continu
Exogenen	continu of categoraal	continu of categoraal	continu of categoraal	continu of categoraal	continu of categoraal	continu of categoraal	continu of categoraal	continu of categoraal
Parametrisch	✓	✓	✓	✓	✓	✓	✓	✓
Toepassingsgebied	bij data zonder bijzonderheden	bij slecht gemeten of onderling gecorreleerde exogenen	bij overfitting	bij uitschieters	bij niet-normaal verdeelde storingsterm	bij afgeknotte variabelen	bij gecensureerde variabelen	bij kleine of scheef verdeelde steekproef
Inhoudelijke uitlegbaarheid van het algoritme	★★★	★★☆	★★☆	★★★	★★★	★★★	★★★	★★☆
Eenvoud van het algoritme	★★★	★★★	★★☆	★☆☆	★★☆	★★☆	★★☆	★★☆
Implementeerbaarheid	★★★	★★★	★★★	★★★	★★☆	★★★	★★★	★★☆
Rekentijd	★★★	★★★	★★★	★★★ (M) ★☆☆ (S,MM)	★★★	★★★	★★★	★☆☆
Domeinkennis	✓	✓	✓	✓	✓	✓	✓	✓
Ketenconsistentie	✓	✓	✓	✓	✓	✓	✓	×
Tijdscomponent mogelijk	✓	✓	✓	✓	✓	×	×	✓
Inhoudelijke uitlegbaarheid van de prognoses	✓	✓	✓	✓	✓	✓	✓	✓
Ruis in de data mogelijk	✓	✓	✓	✓	✓	✓	✓	✓
Privacy: toepasbaar op geaggregeerde data	✓	✓	✓	✓	✓	×	×	✓
indien ×: aggregeerbaarheid						★★★	★★★	
Rechtvaardigheid								verdeling exogenen niet door data bepaald
Overig		lastig om goede instrumenten te vinden				Storingsterm niet normaal verdeeld	Storingsterm niet normaal verdeeld	kansverdeling van de parameters

★☆☆ moeilijk/slecht
★★☆ matig/gemiddeld
★★★ makkelijk/goed

4 Alternatieve specificaties

Zoals gezegd bestaat het huidige PMJ voornamelijk uit lineaire regressies. Maar het is mogelijk om delen van het model op een dusdanig andere wijze te specificeren dat ook een ander algoritme kan worden gebruikt. In paragraaf 4.1 worden een aantal algemene begrippen en definities toegelicht. Paragraaf 4.2 beschrijft het verschil tussen regressie en classificatie. Paragraaf 4.3 gaat in op diverse vormen van lineaire tijdreeksanalyse. Sommige vormen van tijdreeksanalyse worden reeds in het PMJ toegepast. Paragraaf 4.4 gaat in op niet-lineaire regressie methoden. Paragraaf 4.5 behandelt diverse classificatie algoritmes. En in paragraaf 4.6 worden algoritmes behandeld die zowel voor classificatie als regressie kunnen worden ingezet. Ook hier geldt dat het overzicht in dit hoofdstuk geenszins uitputtend is. Er zijn vele algoritmes maar in dit hoofdstuk is geselecteerd op algoritmes die mogelijk relevant kunnen zijn voor het PMJ-model omdat ze een oplossing bieden voor veel voorkomende problemen.

4.1 Terminologie

In deze paragraaf worden een aantal algemene begrippen en definities toegelicht, die in het vervolg van dit rapport veelvuldig worden gebruikt.

4.1.1 Parametrische en niet-parametrische algoritmen

Parametrische algoritmen zijn gebaseerd op een theoretisch model dat de relatie tussen de exogene en de endogene variabelen definieert. Niet-parametrische algoritmen zijn niet gebaseerd op een theoretisch model maar leren van de gegevens zelf. Parametrische algoritmen zijn restrictiever dan niet-parametrische algoritmen omdat de functionele vorm vooraf gekozen wordt, maar dat zorgt er ook voor dat het model sneller en gemakkelijker te trainen is. Parametrische algoritmen hebben niet zoveel trainingsgegevens nodig en kunnen goed werken, zelfs als de aansluiting op de gegevens niet perfect is. Maar parametrische algoritmen zijn wel gevoeliger voor uitschieters. Parametrische algoritmen zijn het meest geschikt voor problemen waarbij de invoergegevens goed gedefinieerd en voorspelbaar zijn. Enkele voorbeelden van parametrische algoritmen zijn:

- (niet-)lineaire regressie (hoofdstuk 3, paragraaf 4.3 en 4.4);
- lineaire discriminantanalyse (paragraaf 4.5.1);
- naïeve Bayes classificatie (paragraaf 4.5.2);
- logistische regressie (paragraaf 4.5.3);
- (simpele) neurale netwerken (paragraaf 4.6.4).

Niet-parametrische algoritmen zijn flexibeler dan parametrische algoritmen. Door geen aannames te doen, zijn ze vrij om elke functionele vorm uit de trainingsgegevens te leren. Hierdoor zijn de resultaten soms wel moeilijker te begrijpen. Ook is er een grotere hoeveelheid gegevens nodig, waardoor ze meer rekentijd nodig hebben. Verder is er een groter risico op overfitting. Niet-parametrische algoritmen zijn het meest geschikt voor problemen waarbij de invoergegevens niet goed gedefinieerd zijn of te complex zijn om te worden gemodelleerd met behulp van een parametrisch algoritme. Enkele voorbeelden van niet-parametrische algoritmen zijn:

- k-nearest neighbours (paragraaf 4.6.1);
- beslisbomen (paragraaf 4.6.2);
- random forest (paragraaf 4.6.2 en 6.2);
- support vector machines met niet-lineaire kernels (paragraaf 4.6.3);
- neurale netwerken met een niet-parametrische activatiefunctie (paragraaf 4.6.4).

4.1.2 Trainingset, validatieset en testset

De beschikbare data voor het maken van een empirisch model wordt meestal gesplitst in drie delen:

- Trainingset, d.w.z. de gegevens die worden gebruikt om de parameters van het theoretisch model te bepalen. Meestal bevat de trainingset 70%-80% van de data.
- Validatieset, d.w.z. de gegevens die worden gebruikt om de zogenoemde hyperparameters af te stemmen. Dit zijn de parameters die worden gebruikt om het leerproces te sturen. De validatieset wordt dus gebruikt om de optimale complexiteit van het empirische model te bepalen. Meestal bevat de validatieset 0%-15% van de data. Sommige algoritmes kennen geen hyperparameters en dan is een validatieset niet nodig.
- Testset, d.w.z. de gegevens die worden gebruikt om de uiteindelijke prestaties van een empirische model te meten met als doel het beste model te kiezen. Meestal bevat de testset 15%-30% van de data.

Om stabielere resultaten te krijgen en alle gegevens voor training te gebruiken, kan een dataset herhaaldelijk worden opgesplitst in meerdere trainings- en validatiedatasets. Dit staat bekend als kruisvalidatie (‘cross validation’). Overigens moeten de gegevens in trainingset, validatieset en testset onder min of meer dezelfde externe omstandigheden zijn gegenereerd. Zo niet, dan is de kans groot dat gewijzigde externe omstandigheden ten onrechte als fout van het model of het algoritme worden aangemerkt.

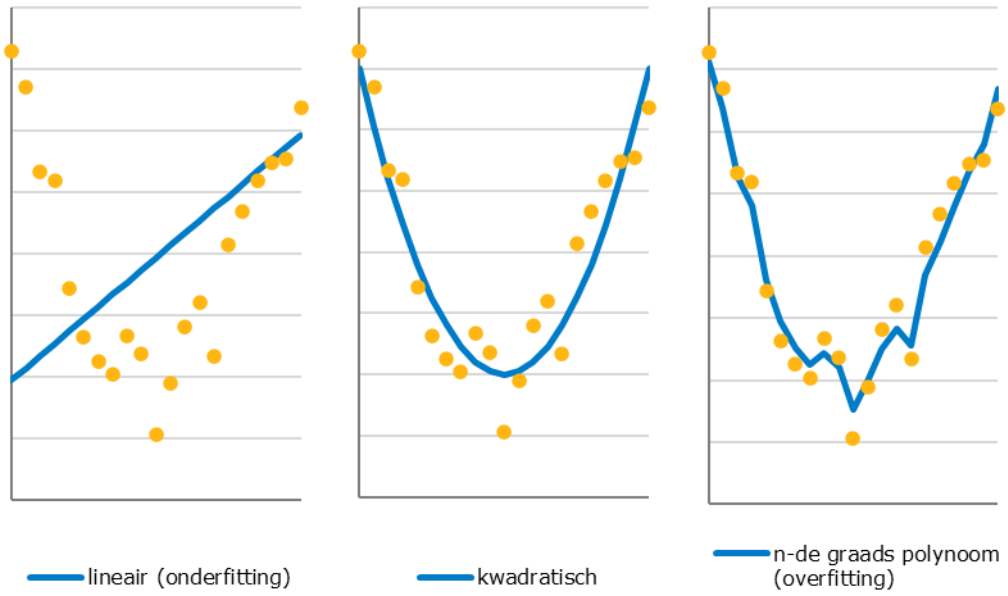
4.1.3 Bias en variantie

De onzuiverheid (‘bias’) en de variantie van de schatter spelen een grote rol in de beoordeling van de kwaliteit van de prognoses van het empirische model. De bias is een vertekening of afwijking van het juiste resultaat die een systematische oorzaak heeft en dus niet te wijten is aan toevallige effecten of aan de steekproef. Hoge bias kan ertoe leiden dat een algoritme de relevante relaties tussen de exogene variabelen en de endogene variabele mist. Dit wordt ook wel ‘underfitting’ genoemd. Bij underfitting is het empirische model te simpel waardoor het niet goed werkt op de trainingset maar ook niet generaliseerbaar is naar nieuwe data.

De variantie is een maatstaf voor de spreiding van een reeks waarden, dat wil zeggen de mate waarin de waarden onderling van elkaar verschillen. Hoe groter de variantie, des te meer de waarden onderling verschillen en dus ook meer van het gemiddelde afwijken. Hoge variantie kan het gevolg zijn van modellering van willekeurige ruis in de trainingset. Dit wordt ook wel ‘overfitting’ genoemd. Als er sprake is van overfitting van het empirische model dan wordt het belang van één of meer exogene variabelen in de relatie met de endogene variabele overschat. Dit maakt het empirische model gevoelig voor kleine veranderingen in de waarden van de exogene variabelen. Bij

overfitting past het empirische model zeer goed op de trainingset maar is het niet generaliseerbaar naar nieuwe data omdat het te specifiek is afgesteld op de beschikbare data in de trainingset (zie figuur 4.1).

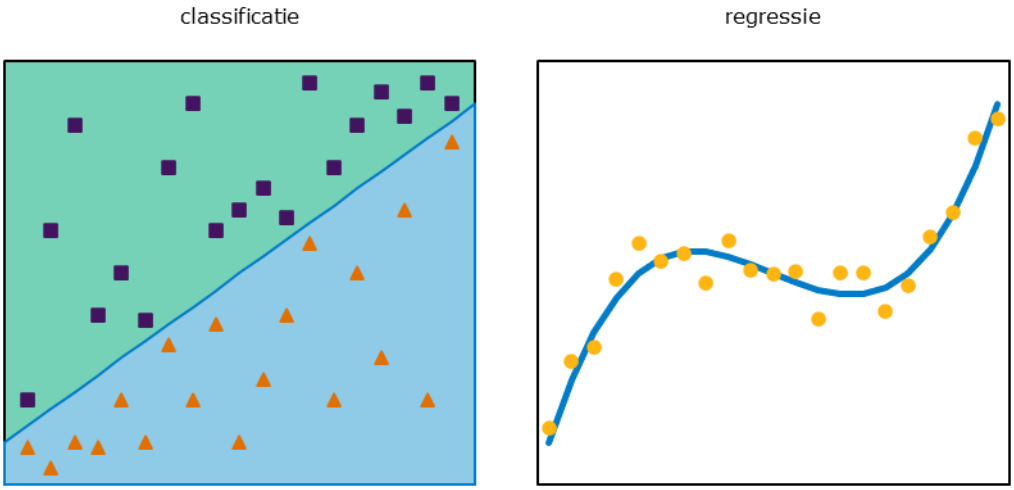
Figuur 4.1 Underfitting en overfitting



4.2 Classificatie versus regressie

Er zijn grofweg drie soorten algoritmes, namelijk classificatie, regressie en clustering. Clustering omvat alle algoritmes die leren zonder toezicht, dat wil zeggen dat er geen endogene variabele is. Dit type algoritme komt in beperkte mate in paragraaf 5.1.2 aan bod. Bij classificatie en regressie is er altijd sprake van leren onder toezicht, dat wil zeggen dat er een endogene variabele is die uiteindelijk moet worden voorspeld. Het verschil tussen regressie en classificatie zit in de waarden die een endogene variabele kan aannemen. Classificatie algoritmes worden gebruikt om de waarde van een categorale endogene variabele te voorspellen. Categorale variabelen zijn variabelen waarvan de waarden geen meetbare hoeveelheden voorstellen en dus niet gesommeerd kunnen worden tot totalen. De waarden zijn in te delen in categorieën, bijvoorbeeld de beslissing van de rechter voor een geldboete, taakstraf of gevangenisstraf. Het aantal categorieën is eindig. Regressie algoritmes worden gebruikt om een waarde van een continue endogene variabele te voorspellen. Continue variabelen hebben waarden waarop berekeningen kunnen worden uitgevoerd, bijvoorbeeld het aantal zaken dat instroomt bij het OM. De waarden die de endogene variabele kan aannemen zijn in principe oneindig. Een regressie algoritme streeft ernaar de meest optimale 'lijn' door de datapunten te trekken, terwijl een classificatie algoritme ernaar streeft om de meeste optimale klassenindeling voor de datapunten te vinden (zie figuur 4.2). Of de exogene variabelen categoraal of continu zijn is niet relevant voor de keuze van het type algoritme.

Figuur 4.2 Classificatie versus regressie



De scheiding tussen deze twee typen algoritmes is overigens niet heel strikt. Zo kan een continue variabele altijd worden gecategoriseerd. Bijvoorbeeld, in plaats van te voorspellen hoeveel zaken per arrondissement bij het OM instromen, kunnen we ook voorspellen of er weinig, gemiddeld of veel zaken instromen per arrondissement. Andersom kan in sommige gevallen een categorale variabele naar een continue variabele worden omgezet. Bijvoorbeeld, in plaats van te voorspellen of iemand een geldboete, taakstraf of gevangenisstraf krijgt opgelegd, kunnen we ook de kans voorspellen dat iemand een geldboete, taakstraf of gevangenisstraf krijgt. De kans kan elke waarde aannemen tussen 0 en 1 en is dus continu. Sommige algoritmes kunnen worden gebruikt voor zowel classificatie als regressie. Ook combinaties zijn mogelijk.

4.3 Lineaire tijdreeksanalyse

In een tijdreeksanalyse zijn prognoses van de endogene variabele gebaseerd op de waarden uit het verleden van de endogene variabele en mogelijk een constante, een trend en seizoensindicatoren. In zijn puurste vorm bevat een tijdreeksanalyse geen exogene variabelen en dus geen domeinkennis, maar het is mogelijk om exogene variabelen toe te voegen. Twee typen modellen worden hier besproken: 'autoregressive integrated moving average' (ARIMA) en 'error correction' modellen (ECM). ARIMA richt zich op de autocorrelaties in de gegevens terwijl ECM als een combinatie van ARIMA en lineaire regressie modellen kan worden beschouwd. Als de theoretische ARIMA- en ECM-modellen niet-stationair zijn, wat betekent dat het gemiddelde en/of de variantie in de loop van de tijd veranderen, moeten deze modellen vóór schatting stationair moeten worden gemaakt.

4.3.1 Autoregressive integrated moving average

Een ARIMA-model gaat ervan uit dat de endogene variabele lineair afhangt van zijn eigen waarden uit het verleden (autoregressie, AR) en van de huidige waarde en waarden uit het verleden van een stochastische term (voortschrijdend gemiddelde, MA) en dat de endogene variabele moet worden gedifferentieerd (integratie, I) om de reeks stationair te maken. Eventueel kunnen exogene variabelen aan het model

worden toegevoegd. Als een ARIMA-model geen MA-termen bevat, is het lineair in de parameters. Als een ARIMA-model na schatting wel MA-termen bevat, is het niet-lineair in de parameters. Maar bij het voorspellen met ARIMA-modellen worden de toekomstige waarden van MA-termen altijd op nul gezet, waardoor ARIMA-voorspelmodellen toch lineair zijn. ARIMA-modellen waarbij de orde van integratie nul is worden ook wel ARMA-modellen genoemd.

Het voordeel van de ARIMA-modellen is dat ze relatief eenvoudig zijn toe te passen, hoewel de toetsen op stationariteit en heteroscedasticiteit (dat wil zeggen niet constante variantie) wel zorgvuldig moeten worden uitgevoerd. Een nadeel is dat stationaire univariate ARIMA-modellen alleen bedoeld zijn voor korte-termijnprognoses, omdat ze weliswaar optimale korte-termijnprognoses opleveren, maar snel terugvallen naar het gemiddelde van het proces (Deadman (2003)). De ARIMA-procedure is ook aangepast voor multivariate analyses (dus meerdere endogene variabelen), bijvoorbeeld Witt en Witte (2000), en voor panelgegevens, in het bijzonder 'space-time' autoregressieve modellen (ST-AR), bijvoorbeeld Shoesmith (2013).

Er zijn verschillende onderzoeken geweest waarbij het ARIMA-model in zijn puurste vorm werd gebruikt om misdaadniveaus te voorspellen. Deadman (2003) paste een ARIMA-model toe op woninginbraken in Engeland en Wales en vergeleek de resultaten met andere typen modellen, waaronder een 'error correction model' (zie volgende paragraaf). Vijayarani et al. (2021) passen ARIMA-modellen toe op geregistreerde misdaden en arrestaties in Chicago, Illinois, VS. Lin et al. (1986) gebruiken een ARIMA-model om de gevangenispopulatie in Louisiana, VS te voorspellen. Goldman et al. (1976) gebruiken een ARIMA-model om de werkdruk voor federale districtsrechtbanken in de VS te voorspellen. Recentelijk hebben Kruisbergen et al. (2024) ARIMA-modellen gebruikt om het effect van COVID-19 op geregistreerde criminaliteit te meten. Op enkele plekken in het huidige PMJ worden ARIMA-modellen al gebruikt, met name in die gevallen waar geen model met exogene variabelen kon worden gevonden of waar in verband met de beperkte beschikbaarheid van data gebruik is gemaakt van maanddata.

In veel onderzoeken worden exogene variabelen aan het ARIMA-model toegevoegd om domeinkennis op te nemen. We spreken dan van een ARIMAX-model. Barros et al. (2020) modelleren criminaliteit in Brazilië met behulp van een ARIMAX-model met exogene variabelen. Wan et al. (2013) gebruiken een ARIMAX-model met exogene variabelen om de gevangenispopulatie in New South Wales, Australië te voorspellen. Recentelijk hebben verschillende studies ARIMAX-modellen gebruikt om het effect van COVID-19 op geregistreerde criminaliteit te meten, bijvoorbeeld Ashby (2020), Payne en Morgan (2020), Payne et al. (2022) en Piquero et al. (2020) en Moolenaar en Choenni (2021). Door een ARIMAX-model te schatten, inclusief correctie voor COVID-19, is het mogelijk om prognoses te maken van geregistreerde criminaliteit alsof de COVID-19-pandemie nooit heeft plaatsgevonden. Door deze prognoses te vergelijken met werkelijke waarden kan het effect van de COVID-19-pandemie op de geregistreerde criminaliteit worden bepaald.

4.3.2 Error Correction Model

Een ECM wordt gebruikt voor het inschatten van zowel korte- als langetermijneffecten van exogene tijdreeksen op een endogene tijdreeks. Een ECM gaat ervan uit dat er een langdurige relatie (soms een evenwicht genoemd) bestaat tussen een endogene

variabele en één of meerdere exogene variabelen, bijvoorbeeld tussen criminaliteit en het werkloosheidspercentage. Wel kunnen er door externe factoren (zoals bijvoorbeeld COVID-19) tijdelijke afwijkingen zijn in de langdurige relatie. Deze afwijkingen kunnen ook de dynamiek van de korte termijn relatie beïnvloeden. Daarom bevat de vergelijking voor de korte termijn relatie een correctie voor afwijkingen van de lange termijnrelatie in het verleden. Om een ECM te implementeren moeten alle tijdreeksen in het model eerst worden getest op heteroscedasticiteit en niet-stationariteit.

De (impliciete) onderliggende aanname bij ECM is dat er daadwerkelijk een langdurige relatie bestaat. In dit geval is een ECM geschikter dan andere tijdreeksmethoden. De vraag rijst of dit type model geschikt is als er geen sprake is van een dergelijke langdurige relatie of als de relatie instabiel is. Hendry en Clements (1998) merkten op dat prognoses van een ECM slecht kunnen zijn wanneer er een verschuiving is in de langetermijnrelatie. Deadman (2003) vergeleek resultaten van ARIMA- en ECM-modellen voor woninginbraken in Engeland en Wales in 1998-2001 en concludeerde dat er geen langetermijnrelatie bestaat, gebaseerd op het feit dat ARIMA-modellen betere prognoses opleverden dan ECM-modellen. Andere voorbeelden van ECM toegepast op criminaliteit zijn Dhiri et al. (1999), Virén (2010) en Harries (2003). De ECM-procedure is ook aangepast om meerdere endogene variabelen in één model ('vector error correction model', VECM) te accommoderen, bijvoorbeeld Saridakis (2004). In het huidige PMJ worden ECM's reeds gebruikt om een deel van de geregistreerde misdrijven te voorspellen.

4.3.3 Samenvatting en implicaties voor het PMJ

Een tijdreeksanalyse algoritmes zoals ARIMA wordt veelvuldig toegepast, omdat het relatief makkelijk is te implementeren en in veel softwarepakketten standaard is geïntegreerd. Daarin ligt ook de valkuil: voor correcte toepassing van ARIMA en ECM, die beide ook in het huidige PMJ worden toegepast, moet vooraf gecontroleerd worden of de data aan bepaalde voorwaarden voldoet. Een specifiek nadeel van ECM is dat het bestaan van een lange termijn relatie wordt verondersteld. Een nadeel is dat tijdreeksanalyses vooral geschikt zijn voor korte-termijnprognoses, omdat de modellen op de lange termijn de neiging hebben terug te keren naar het gemiddelde van het proces. Desalniettemin kan met name ARIMA een interessante bijdrage vormen voor het PMJ. De eigenschappen van de diverse algoritmes zijn in tabel 4.1 nog eens op een rijtje gezet.

Tabel 4.1 Kenmerken van niet-lineaire regressie-algoritmes

	ARIMA(X)	ECM
Algoritme	regressie	regressie
Endogene	continu	continu
Exogenen	geen of continu of categoraal	continu of categoraal
Parametrisch	✓	✓
Toepassingsgebied	tijdreeksen	tijdreeksen
Inhoudelijke uitlegbaarheid van het algoritme	★★★	★★★
Eenvoud van het algoritme	★★★	★★★
Implementeerbaarheid	★★★	★★★
Rekentijd	★★★	★★★
Domeinkennis	✓ (met exogenen) × (zonder exogenen)	✓
Ketenconsistentie	✓ (met exogenen) × (zonder exogenen)	✓
Tijdscomponent mogelijk	✓	✓
Inhoudelijke uitlegbaarheid van de prognoses	✓	✓
Ruis in de data mogelijk	✓	✓
Privacy: toepasbaar op geaggregeerde data	✓	✓
indien ×: aggregeerbaarheid		
Rechtvaardigheid issues		
Overig	Zonder MA is het model lineair in de parameters.	Aanname dat er een lange termijn relatie bestaat; zonder MA is het model lineair in de parameters.

★★★ moeilijk/slecht
★★★ matig/gemiddeld
★★★ makkelijk/goed

4.4 Niet-lineaire regressie

In een niet-lineaire regressie wordt de relatie tussen de endogene variabele en de exogene variabele(n) beschreven met een theoretisch model dat niet-lineaire parameters bevat, bijvoorbeeld de vermenigvuldiging van twee parameters. In de praktijk wordt dit niet veel toegepast omdat het vaak tot convergentieproblemen van het algoritme leidt. Soms is het wel mogelijk om een niet-lineair theoretisch model te transformeren naar een lineair theoretisch model, bijvoorbeeld door de natuurlijk logaritme van het hele model te nemen. Vervolgens kan dat lineaire regressie zoals beschreven in hoofdstuk 3 op de transformatie worden toegepast. Hieronder worden vier specifieke vormen van niet-lineaire regressie beschreven.

4.4.1 Tijdreeksanalyse middels exponential smoothing

Een niet-lineaire vorm van tijdreeksanalyse is 'exponential smoothing' ('Error-Trend-Seasonality' ofwel ETS). ETS richt zich meer op de trend en seizoenscomponenten in de gegevens, terwijl ARIMA (zie paragraaf 4.3.1) zich richt op de autocorrelaties in de gegevens. Alle ETS-modellen zijn niet-stationair. Er is een kleine overlap tussen ETS- en ARIMA-modellen (voor details zie Hyndman & Athanasopoulos, 2021).

Exponential smoothing is een algoritme voor het maken van adaptieve prognoses. De prognose van de endogene variabele is een gewogen gemiddelde van de waarden uit het verleden met exponentieel afnemende gewichten in de loop van de tijd. Eenvoudige theoretische smoothing-modellen bestaan al vele jaren en bevatten een trendcomponent (T) en een seizoenscomponent (S). Bij exponential smoothing wordt een derde component toegevoegd: de foutterm (E). De drie componenten kunnen in verschillende combinaties worden gecombineerd om de endogene variabele te schatten, maar altijd in dezelfde volgorde. De trendcomponent kan worden ontleed in een niveauterm, die altijd aanwezig is, en een groeiterm. De groeitermijn kan worden gedempt, waardoor de impact van de trend in de tijd afneemt. Er zijn veel verschillende vormen van ETS; zie Gardner (2006) voor een uitgebreide behandeling van ETS. Chua & Tumibay (2020) pasten deze methode toe op misdaadrapportages in Angeles City op de Filippijnen. Gorr et al. (2003) pasten deze methode toe op criminaliteit in Pittsburgh, PA, VS.

Een groot voordeel van ETS ten opzichte van eenvoudige smoothing-modellen is dat informatiecriteria gebaseerd op de waarde van de aannemelijkheidsfunctie en het aantal parameters kunnen worden gebruikt voor modelselectie. Een ander voordeel van ETS is dat er niet veel waarnemingen nodig zijn om een prognose te maken. Tegelijkertijd is dit ook een nadeel. Als het aantal waarnemingen beperkt is en niet representatief is voor een 'normale' situatie, is het onwaarschijnlijk dat deze methode nauwkeurige prognoses zal opleveren. Een ander nadeel is dat het onderscheid tussen trendcomponent, seizoenscomponent en foutterm kunstmatig is, waardoor de resultaten moeilijk interpreteerbaar zijn.

4.4.2 Algoritmes voor aftelbare gegevens

In het geval van aftelbare gegevens ('count data') kan een Poisson-proces worden gebruikt om te voorspellen. Een Poisson-model is een model voor een reeks afzonderlijke gebeurtenissen, bijvoorbeeld het dagelijks aantal slachtoffers dat naar

het politiebureau komt om aangifte te doen. Aangenomen wordt dat de exacte timing van deze gebeurtenissen willekeurig is, maar de gemiddelde tijd tussen gebeurtenissen bekend is. Als er bijvoorbeeld 180 misdrijven per jaar worden gemeld, is de gemiddelde tijd tussen aangiften ongeveer twee dagen. Er kunnen echter dagen zijn dat er geen slachtoffers binnenkomen of er kunnen tien opeenvolgende dagen zijn met één of meer slachtoffers per dag. De aankomst van een gebeurtenis moet onafhankelijk zijn van de vorige gebeurtenis en de gemiddelde aankomsttijd moet constant zijn. Aangenomen wordt dat de aankomsten een Poisson-verdeling volgen, maar het is ook mogelijk om een negatieve binomiale, exponentiële of normale verdeling te specificeren.

Een Poisson-model werkt goed als gebeurtenissen willekeurig plaatsvinden en er geen informatie beschikbaar is over aankomsttijden en de data veel nullen of kleine aantallen bevat. In plaats van gebeurtenissen in de tijd is het ook mogelijk om te kijken naar gebeurtenissen in een gebied. Om bijvoorbeeld de relaties tussen misdrijven en sociaal-economische determinanten te onderzoeken, formuleren Hu et al. (2018) een space-time model op basis van een binominale verdeling en een Poisson-verdeling en vergelijken de resultaten. Liu en Brown (2003) voorspellen criminele incidenten met behulp van een overgangsdichtheidsmodel voor prognose van gebeurtenissen in de tijd en ruimte.

4.4.3 Algoritmes voor duurgegevens

Duurgegevens ('duration' data) kunnen worden geanalyseerd door middel van een 'survival' analyse. Het doel is de verwachte tijd vanaf het begin van een bepaalde gebeurtenis tot aan het eind van deze gebeurtenis dan wel het einde van het meetmoment te bepalen. Het begin van de gebeurtenis kan per waarneming verschillen. De endogene variabele bestaat uit twee delen: tijd tot aan het einde van een gebeurtenis en een indicator die aangeeft of de gebeurtenis wel of niet heeft plaatsgevonden binnen de meetperiode (zie box 4.1 voor een voorbeeld). Aan het einde van de meetperiode zal voor een deel van de waarnemingen de gebeurtenis waarin we zijn geïnteresseerd, nog niet hebben plaatsgevonden. Dit concept staat bekend als censureren en is een van de belangrijkste uitdagingen van survivalanalyse. In het geval van (semi-)parametrische survivalanalyse is de wijze waarop hiermee wordt omgegaan hetzelfde als bij Tobit-regressie (zie paragraaf 3.1.6).

Box 4.1 Voorbeelden van de analyse van duurgegevens

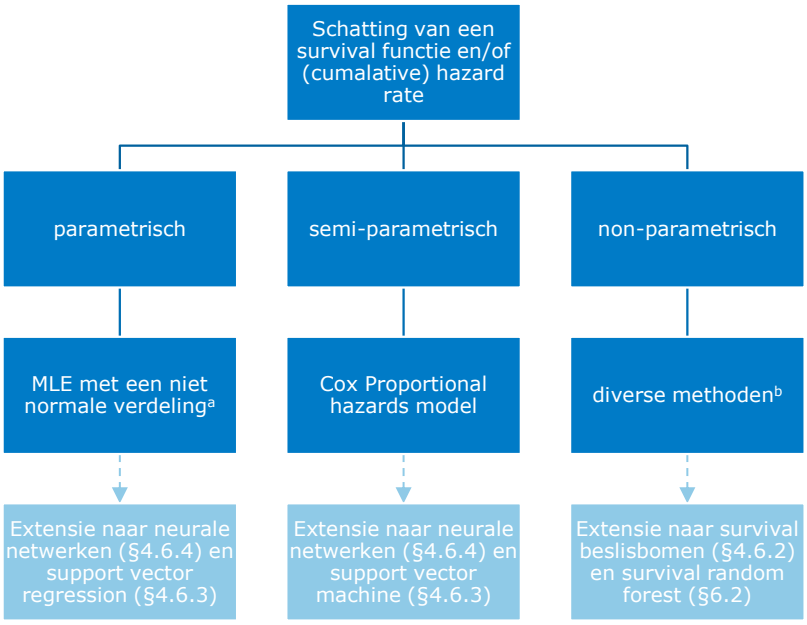
De tbs-maatregel kent geen vaste einddatum in tegenstelling tot een gevangenisstraf. Wanneer een tbs-er binnenkomt in een tbs-inrichting, is dus vooraf niet bekend wanneer deze persoon weer zal uitstromen. Door middel van een duurmodel kan worden geschat wat de verwachte verblijfstijd zal zijn van het moment van binnenkomst. Als het model wordt getraind met gegevens tot met een bepaalde datum, zal in de dataverzameling de uitstroomdatum van tbs-ers die na deze datum uitstromen, ontbreken. Deze waarnemingen zijn dan gecensureerd. Andere mogelijk toepassingsgebieden zijn doorlooptijden of de duur van de voorlopige hechtenis.

Bij survivalanalyse zijn we geïnteresseerd in de kans dat de tijd tot aan een gebeurtenis langer of gelijk is aan een bepaalde tijd. Dit wordt vastgelegd door de zogenoemde 'survival'-functie (overlevingsfunctie). De survivalfunctie is afhankelijk van de zogenoemde 'hazard rate' (risico). De hazard rate geeft aan wat de kans is dat

een gebeurtenis plaatsvindt op een bepaald tijdstip, gegeven het feit dat het nog niet heeft plaatsgevonden. In het meest simpele geval is de hazard rate een constante. Maar de hazard rate kan ook een functie zijn van exogene variabelen. De 'cumulative hazard rate' combineert de schattingen van de hazard rate over alle tijdstippen.

Er zijn verschillende methoden om de survival-functie en de (cumulative) hazard rate te schatten (zie figuur 4.3). De Kaplan-Meier-schatter schat de survival-functie op niet-parametrische wijze. De Nelson-Aalen-schatter geeft een niet-parametrische schatting van de cumulatieve hazardfunctie. Bij beide schatters is het niet mogelijk om exogene variabelen mee te nemen in de analyse. Het Cox 'proportional hazards'-model is een semi-parametrisch lineair regressiemodel dat een schatting maakt van de hazardfunctie op basis van exogene variabelen. Het is semiparametrisch omdat er geen aanname wordt gemaakt over de vorm van de hazardfunctie. Het is vergelijkbaar met lineaire regressie (zie paragraaf 3.1) en logistische regressie (zie paragraaf 4.5.3). De MLE is een parametrische aanpak waarbij een verdelingsfunctie wordt gekozen die geschikt is voor dit type gecensureerde data, zoals de exponentiele, Weibull of lognormale verdeling. De genoemde methoden kunnen worden uitgebreid naar beslisbomen, 'support vector machines/regression' en neutrale netwerken voor survivalanalyse (zie paragraaf 4.6). Dit rapport gaat hierop niet specifiek in.

Figuur 4.3 Voorbeelden van methoden voor een survivalanalyse



a bijvoorbeeld exponentiele, Weibull, lognormale, gamma of log-logistische verdeling.

b bijvoorbeeld 'life table', Kaplan-Meier schatter of Nelson-Aalen schatter.

4.4.4 Samenvatting en implicaties voor het PMJ

Een tijdreeksanalyse algoritme zoals ETS wordt veelvuldig toegepast, omdat het relatief makkelijk is te implementeren en in veel softwarepakketten standaard is geïntegreerd. Maar de prognoses van ETS zijn lastig interpreteerbaar. Een ander

nadeel is dat tijdreeksanalyses vooral geschikt zijn voor korte-termijnprognoses, omdat de modellen op de lange termijn de neiging hebben terug te keren naar het gemiddelde van het proces. Andere vormen van niet-lineaire regressie die mogelijk interessant kunnen zijn voor het PMJ zijn algoritmes voor duurgegevens. Met name een (semi-)parametrische survivalanalyse van duurgegevens lijkt veelbelovend voor die onderdelen van MinJenV waarvoor de duur niet vooraf bekend is, zoals de tbs- of pij-maatregel of voorlopige hechtenis. De eigenschappen van de diverse algoritmes zijn in tabel 4.2 nog eens op een rijtje gezet.

Tabel 4.2 Kenmerken van niet-lineaire regressie-algoritmes

	ETS	Poisson	Survival (semi-) parametrisch	Survival non-parametrisch
Algoritme	regressie	regressie	regressie	regressie
Endogene	continu	continu	continu	continu
Exogenen	n.v.t.	continu of categoraal	Geen of continu of categoraal	n.v.t.
Parametrisch	✓	✓	✓	×
Toepassingsgebied	tijdreeksen	veel nullen en kleine waarden in de endogene	duurdata	duurdata
Inhoudelijke uitlegbaarheid van het algoritme	★☆☆	★★★	★★★	★★★
Eenvoud van het algoritme	★★☆	★★★	★★★	★★★
Implementeerbaarheid	★★★	★★★	★★☆	★★☆
Rekentijd	★★★	★★★	★★★	★★★
Domeinkennis	×	✓	✓ (met exogenen) × (zonder exogenen)	×
Ketenconsistentie	×	✓	✓ (met exogenen) × (zonder exogenen)	×
Tijdscomponent mogelijk	✓	✓	×	×
Inhoudelijke uitlegbaarheid van de prognoses	n.v.t.	✓	✓	n.v.t.
Ruis in de data mogelijk	✓	✓	✓	×
Privacy: toepasbaar op geaggregeerde data	✓	✓	× uitkomsten hebben geen betrekking op individuen	× uitkomsten hebben geen betrekking op individuen
indien ×: aggregeerbaarheid			★★☆	★★☆
Rechtvaardigheid issues				
Overig		Aanname: onafhankelijkheid van aankomsten		

★☆☆ moeilijk/slecht
★★☆ matig/gemiddeld
★★★ makkelijk/goed

4.5 Classificatie

Deze paragraaf beschrijft een aantal technieken die uitsluitend geschikt zijn voor classificatie, dus waarbij de endogene variabele categoraal is. Er zijn grofweg twee benaderingen: generatief model en discriminatief model. Een generatief model gaat uit van een gezamenlijke kansverdeling van de endogene en exogene variabelen. Een discriminatief model gaat uit van óf de voorwaardelijke kansverdeling van de endogene variabele gegeven de exogenen óf helemaal geen kansverdeling. Voorbeelden van een generatief model zijn lineaire discriminantanalyse en naïeve Bayes-classificatie (zie respectievelijk paragraaf 4.5.1 en 4.5.2). Voorbeelden van een discriminatief model zijn logistische regressie en support vector machine (zie paragraaf 4.5.3 en 4.6.3).[4]

4.5.1 Discriminantenanalyse

Discriminantenanalyse is een techniek die wordt gebruikt als de endogene variabele categoraal is en de exogene variabelen continu. Het doel is om een combinatie van de exogene variabelen te vinden die de categorieën van de endogene variabele maximaal van elkaar scheidt. Er zijn een aantal varianten zoals lineaire discriminantenanalyse, Fishers discriminantenanalyse en Kwadratische discriminantenanalyse.

Om lineaire discriminantenanalyse (LDA) toe te passen moeten de data aan een aantal voorwaarden voldoen:

- De gegevens zijn lineair scheidbaar.
- De exogene variabelen zijn normaal verdeeld voor elke categorie van de endogene variabele.
- De (co)variantie is hetzelfde voor elke categorie van de endogene variabele (homoscedasticiteit).
- Er is weinige onderlinge correlatie tussen de exogene variabelen (lage multicollineariteit).
- De waarnemingen zijn onafhankelijk.

Als aan alle bovengenoemde voorwaarden wordt voldaan is LDA een heel goed classificatie algoritme, maar in de praktijk wordt hieraan heel vaak niet voldaan. Indien de (co)variantie niet gelijk is voor elke categorie van de endogene variabele maar wel aan alle andere voorwaarden is voldaan (m.u.v. lineaire scheidbaarheid), kan kwadratische discriminantenanalyse (QDA) worden gebruikt. Dit is een meer algemene versie van LDA. Een andere optie is Fishers discriminanten analyse (FDA). FDA is vergelijkbaar met LDA, dus ook lineair, maar zonder de aannames dat de exogene variabelen normaal verdeeld zijn en dat de (co)variantie hetzelfde is voor elke categorie van de endogene variabele.

[4] Een discriminatief model of een discrimantenanalyse is iets anders dan een discriminerend algoritme (of model of analyse). De laatste jaren is er veel aandacht voor discriminerende algoritmes, dat wil zeggen dat het algoritme (indirect) ten onrechte bepaalde bevolkingsgroepen benadeelt op basis van bepaalde persoonskenmerken. De termen discriminatief model en discrimantenanalyse zijn echter technische termen, waarbij de eerste betrekking heeft op de kansverdeling van de variabelen en de tweede op de zogenoemde discriminantenfunctie, dat wil zeggen een lcombinatie van voorspellers waardoor een nieuwe latente variabele ontstaat. Merk overigens op dat in essentie het doel van alle algoritmes die hier worden behandeld, is om de data in verschillende categorieën onder te verdelen zonder aan deze categorieën een waarde-oordeel te verbinden.

LDA is een eenvoudig en rekenkundig efficiënt algoritme. LDA is geschikt voor kleine steekproeven. Maar de voorgenoemde eisen aan de data vormen wel een beperking. Een groot aantal exogene variabelen kan ook een probleem zijn. LDA is vrij gevoelig voor uitschieters en de omvang van de kleinste categorie moet groter zijn dan het aantal exogene variabelen.

4.5.2 Naïeve Bayes Classificatie

Naïeve Bayes Classificatie is een familie van probabilistische algoritmes gebaseerd op de stelling van Bayes[5] met de 'naïeve' aanname dat de exogene variabelen onafhankelijk zijn en in gelijke mate bij dragen aan de uitkomst. Voor elke waarde van de endogene variabele wordt een kansverdeling geschat volgens de stelling van Bayes gegeven de bijbehorende waarden van de exogene variabelen. Met deze kansverdeling kunnen de nieuwe waarden van de endogene variabele worden geclassificeerd op basis van de bijbehorende waarden van de exogene variabelen. De uitkomst van het naïeve Bayes algoritme is een score. Het wordt pas een classificatie in combinatie met een beslisregel.

Het belangrijkste verschil met frequentistische statistiek is, dat wordt aangenomen dat de exogene variabelen uit een kansverdeling voortkomen, waarbij statistische grootheden zoals het gemiddelde en de variantie van de trainingset de parameters van de verdelingsfunctie bepalen. Voor continue exogene variabelen wordt vaak de normaalverdeling gebruikt en voor categorale exogene variabelen een multinomiale verdeling. Maar als de trainingset heel groot is, dan is er vrijwel geen verschil tussen de frequentistische en de bayesiaanse aanpak. De naïeve Bayes-classificator met exogenen die normaal verdeeld zijn is equivalent aan de kwadratische discriminantenanalyse met diagonale covariantiematrices.

Hoewel dit algoritme technisch gezien niet moeilijk is, zijn de resultaten minder intuïtief als men geen kennis van of affiniteit met kansverdelingen heeft. Vanwege de naïeve aanname is het lastiger om de ketenconsistentie in het algoritme te brengen. Omdat de vorm van de kansverdeling van de exogene variabelen vooraf wordt gekozen en niet op basis van de data wordt bepaald, wordt het risico op een discriminerend algoritme verminderd.

4.5.3 Logistische regressie

Logistische regressie (ook wel logit regressie genoemd) is een veelgebruikt algoritme voor classificatie. Het tussendoel is het voorspellen van de kans dat de waarde van een endogene variabele tot een bepaalde categorie behoort. Hierbij wordt aangenomen dat deze kans een logistische verdeling volgt. Ondanks dat logistische regressie een classificatiemethode voor categorale endogene variabele is wordt deze methode regressie genoemd omdat het tussendoel, namelijk de kans dat de endogene variabele tot een categorie behoort, continu is. Logistische regressie wordt pas een classificatiemethode als er een beslisregel (ook wel drempelwaarde genoemd) wordt ingesteld.

Er zijn verschillende vormen van logistische regressie. We noemen hier de drie vormen die het meest gebruikt worden in de praktijk. De meest eenvoudige is de binomiale logistische regressie. De endogene variabele kan dan slecht twee waarden aannemen:

[5] De stelling van Bayes drukt de kans op A gegeven B uit in de voorwaardelijke kansen op B bij elke mogelijke waarde van A.

'0' of '1', 'geslaagde taakstraf' of 'mislukte taakstraf', gevangenisstraf of geen gevangenisstraf, enz. Als de endogene variabele meerdere categorieën kent dan biedt de multinomiale logistische regressie uitkomst, bijvoorbeeld wanneer de rechter moet beslissen tussen boete, taakstraf of gevangenisstraf. Andere benamingen van multinomiale logistische regressie zijn 'polytomous logistic regression', 'multiclass logistic regression', 'softmax regression', 'maximum entropy classifier', en 'conditional maximum entropy model'. Merk op dat de binomiale logistische regressie een verbijzondering is van de multinomiale logistische regressie en dat naïeve Bayes algoritme toegepast op binaire exogene variabelen gelijk is aan multinomiale logistische regressie. Maar soms hebben de categorieën van de endogene variabele een logische volgorde, bijvoorbeeld boetes van de eerste t/m zesde categorie. In dat geval kan een geordende logistische regressie worden gebruikt.

Een logistische regressie gaat uit van een logistische verdeling, dat wil zeggen een verdeling met een liggende sterk uitgerekte S-vorm. In plaats van de logistische verdeling kan ook een normaalverdeling worden gebruikt. We spreken dan van een probit regressie. De logistische en de normale verdeling lijken veel op elkaar. Het voornaamste verschil is dat de logistische verdeling dikkere staarten heeft. Welke verdeling gebruikt moet worden, hangt dus af van de data.

Logistische regressie heeft een lage bias en een hoge variantie, terwijl naïeve Bayes een hoge bias en lage variantie heeft. Met naïeve Bayes is het makkelijker om te voorspellen met een klein aantal exogene variabelen en weinig observaties en beperkt aantal trainingsets. Net als bij andere regressie algoritmes is wordt bij logistische regressie niet aangenomen dat de exogene variabelen statistisch onafhankelijk van elkaar zijn, in tegenstelling tot bijvoorbeeld een naïeve Bayes-classificator. De multicollineariteit wordt wel verondersteld relatief laag te zijn, omdat het anders lastig wordt om onderscheid te maken tussen de impact van verschillende exogene variabelen. Dit maakt het trainen van een model met logistische regressie model wel langzamer dan met een naïeve Bayes-classificator. Daarom is logistische regressie minder geschikt voor het voorspellen van een endogene variabele met een groot aantal categorieën.

Logistische regressie lijkt sterk op discriminantenanalyse en beide kunnen worden gebruikt om dezelfde onderzoeksvragen te beantwoorden. Maar logistische regressie heeft minder aannames en beperkingen dan discriminantenanalyse, waardoor in de praktijk de voorkeur wordt gegeven aan logistische regressie, omdat zelden wordt voldaan aan de aannames van discriminantenanalyse. Net als bij discriminantenanalyse kan logistische regressie bij onzorgvuldig gekozen exogene variabelen leiden tot ongewenste (d.w.z. discriminerende) categorieën. Ketenconsistentie inbrengen in het algoritme is lastig. Weliswaar kunnen de producten of uitstroom van voorgaande ketenpartners als exogene variabelen worden meegenomen in het algoritme maar omdat de uitkomsten kansen zijn en geen aantallen, kunnen deze resultaten moeilijker worden doorgegeven naar de volgende ketenpartner. Ook is het lastig om de tijdscomponent in te bouwen die nodig is om een langere periode vooruit te voorspellen. De volgende subparagrafen gaan verder in op nog een aantal specifieke problemen die zich vaak voor doen bij logistische regressie.

Propagatie van voorspelfouten
Het grote voordeel van de multinomiale logistische regressie is dat de kansen van alle categorieën met één algoritme worden bepaald en dat er dus sprake is van één voorspelfout. In principe zou een multinomiale logistische regressie kunnen worden

opgesplitst in meerdere binomiale logistische regressies (of andere binaire keuzemodellen). Maar bij elke prognose die wordt gemaakt, kan er sprake zijn van een voorspelfout. Deze fouten hebben de neiging zich te vermenigvuldigen. Dit probleem staat bekend als propagatie van voorspelfouten ('error propagation') en is een serieus probleem in een geschakelde reeks van binaire keuze modellen (zie voorbeeld in box 4.2). Het voorspellen van de kans op elke mogelijke uitkomst in één algoritme, in plaats van reeks van uitkomsten met behulp van meerdere toepassingen van een algoritme, is een manier om dit probleem op te lossen.

Box 4.2 Voorbeeld van propagatie van voorspelfouten

De keuze van de rechter tussen boete, taakstraf of gevangenisstraf kan worden opgesplitst in een eerste keuze tussen wel of niet een gevangenisstraf, en, in het geval dat er niet voor een gevangenisstraf wordt gekozen, in een tweede keuze tussen boete of taakstraf. Stel dat de nauwkeurigheid van de voorspelling van de eerste keuze 90% is en van de tweede keuze 80%. Dan is algemene nauwkeurigheid van het model slechts 72% (=0.9*0.8=0.72).

Onafhankelijkheid van irrelevante alternatieven

Het onderliggende uitgangspunt van multinomiale logistische regressie is dat er sprake is van onafhankelijkheid van irrelevante alternatieven (zie box 4.3 voor een voorbeeld). Dit axioma stelt dat de kans om categorie A boven categorie B te verkiezen alleen afhangt van individuele voorkeuren voor A en B en niet van de aan- of afwezigheid van andere, irrelevante alternatieven (Arrow, 1963). In de praktijk wordt dit axioma vaak geschonden omdat het axioma geen rekening houdt met het feit dat sommige alternatieven perfecte substituten van elkaar zijn. Als het doel van de analyse is om te voorspellen hoe keuzes zouden veranderen als één alternatief zou verdwijnen (bijvoorbeeld als een politieke kandidaat zich terugtrekt uit een race met drie kandidaten), dan is geneste logistische regressie of de multinomiale probit modellen te prefereren omdat bij deze algoritmes schending van het axioma van onafhankelijkheid van irrelevante alternatieven is toegestaan. Dit rapport gaat niet nader in op deze algoritmes.

Box 44.3 Voorbeeld van onafhankelijkheid van irrelevante alternatieven

Een bekend voorbeeld van de schending van het axioma van onafhankelijkheid van irrelevante alternatieven komt van McFadden (1974). Een forens heeft de keuze om met de auto of een rode bus naar het werk te gaan. Deze persoon heeft geen voorkeur, dus de kans voor beide alternatieven is 50% en de verhoudingen tussen beide kansen is dus één op één. Stel er wordt een derde alternatief aangeboden, namelijk een blauwe bus. Ervanuit gaande dat de kleur van de bus volstrekt onbelangrijk is voor de forens, zou men verwachten dat de kans op de auto 50% blijft, terwijl de kans dat de forens met één van de bussen gaat 25% is. Volgens de hypothese van onafhankelijkheid van irrelevante alternatieven is dat niet toegestaan, want de kansverhouding tussen rode bus en auto mag niet veranderen. Omdat de forens ook geen voorkeur heeft voor de kleur van de bus, impliceert dit dat de kansen van de rode en blauwe bus gelijk zijn. Dus de nieuwe kansen worden: 33% voor de auto, 33% voor de rode bus en 33% voor de blauwe bus. De kans om met de auto te gaan is dus gedaald van 50% naar 33% door de introductie van een blauwe bus. Dit is volstrekt onlogisch. Het probleem met het axioma van onafhankelijkheid van irrelevante alternatieven is dat het geen rekening houdt met het feit dat de rode bus en de blauwe bus perfecte substituten zijn.

4.5.4 Beslisregels

Logistische regressie en naïeve Bayes worden pas classificatiemethoden als er een beslisregel wordt ingesteld. Het tussendoel is het voorspellen van de kans dat de waarde van een endogene variabele tot een bepaalde categorie behoort, bijvoorbeeld de kans dat een verdachte een gevangenisstraf krijgt opgelegd. Met behulp van een beslisregel op de geschatte kans wordt het einddoel bereikt, namelijk voorspellen tot welke categorie de waarde van de endogene variabele behoort. Bijvoorbeeld, als de geschatte kans groter dan of gelijk aan 50% is dan nemen we aan dat de waarde een gevangenisstraf is, maar als de geschatte kans kleiner dan 50% is, dan nemen we aan dat de waarde geen gevangenisstraf is. De beslisregel kan afhankelijk zijn van het classificatieprobleem zelf en wordt in grote mate beïnvloed door de mate waarin men foutieve prognoses wil voorkomen (zie ook paragraaf 7.2.1).

4.5.5 Samenvatting en implicaties voor het PMJ

Discriminantenanalyse is eenvoudig en een rekenkundig efficiënt algoritme, maar vooral geschikt voor kleine steekproeven met een beperkt aantal exogene variabelen. In het geval van lineaire discriminantenanalyse vormen de gestelde eisen aan de data een beperking en maakt dit algoritme voor het PMJ minder bruikbaar. Omdat de eisen aan de data bij kwadratische discriminanten analyse minder stringent zijn, is dit wel een interessante optie. Ook het naïeve Bayes algoritme is technisch gezien niet moeilijk, maar de resultaten zijn minder intuïtief als men geen kennis van of affiniteit met kansverdelingen heeft. Voor al deze algoritmes geldt dat het inbouwen van ketenconsistentie en de tijdscomponent lastig is.

Voor PMJ lijkt logistische regressie een logische optie: binnen de justitiële ketens zijn veel keuzemomenten. Moet een verdachte preventief worden gehecht? Moet een verdachte wel of niet worden vervolgd en/of berecht? Welke type sanctie moet worden opgelegd? Dit zijn typisch keuzes die met een logistische regressie kunnen worden voorspeld. Desalniettemin zijn er ook beperkingen. Het doel van het PMJ is om zeven jaar vooruit te voorspellen. Het is lastig om het tijdsaspect in een logistische regressie

op te nemen. De enige optie is in de meeste gevallen om de gevonden kansen voor de hele voorspelperiode constant te veronderstellen. Bovendien is voor een dergelijke analyse microdata of (balanced) panel data nodig. Deze is in potentie beschikbaar, maar daaraan zijn wel privacy issues verbonden. De vraag is of de extra werkzaamheden die dat met zich meebrengt opweegt tegen een vergroot inzicht in de prognoses. De eigenschappen van de diverse algoritmes zijn in tabel 4.3 nog eens op een rijtje gezet.

Tabel 4.3 Kenmerken van classificatie-algoritmes

	Discrimantenanalyse	Naïeve Bayes	Logistische regressie
Algoritme	classificatie	classificatie	classificatie
Endogene	categoraal	categoraal	categoraal
Exogenen	continu	continu of categoraal	continu of categoraal
Parametrisch	✓	✓	✓
Toepassingsgebied	Bij weinig exogenen	Bij kleine steekproeven; Berekening van kansen	Berekening van kansen
Inhoudelijke uitlegbaarheid van het algoritme	★★★	★★★	★★★
Eenvoud van het algoritme	★★★	★★★	★★★
Implementeerbaarheid	★★★	★★★	★★★
Rekentijd	★★★	★★★	★★★
Domeinkennis	✓	✓	✓
Ketenconsistentie	×	× (kans) ✓ (conversie naar aantallen)	× (kans) ✓ (conversie naar aantallen)
Tijdscomponent mogelijk	×	×	×
Inhoudelijke uitlegbaarheid van de prognoses	×	✓ (kans) × (conversie naar aantallen)	✓ (kans) × (conversie naar aantallen)
Ruis in de data mogelijk	×	×	✓
Privacy: toepasbaar op geaggregeerde data	×	×	×
indien ×: aggregeerbaarheid	★★★	★★★	★★★
Rechtvaardigheid issues	Mogelijk ongewenste categorieën	Voordeel: verdeling niet door data bepaald	Mogelijk ongewenste categorieën
Overig	Nadeel: endogene variabele mag slechts twee categorieën hebben; De data moet aan veel voorwaarden voldoen	hoge bias en lage variantie	lage bias en een hoge variantie

★★★ moeilijk/slecht
★★★ matig/gemiddeld
★★★ makkelijk/goed

4.6 Algoritmes voor regressie en classificatie

In deze paragraaf worden een aantal algoritmes besproken die zowel voor regressie (continue endogene variabele) als classificatie (categorale endogene variabele) kunnen worden gebruikt. De meeste van deze algoritmes zijn non-parametrisch dus de vorm van het model is niet van tevoren vastgelegd maar wordt door de data bepaald.

4.6.1 K-nearest neighbours

Het 'k-nearest neighbours' algoritme (KNN) gebruikt het concept van afstand om te voorspellen. In het meest simpele geval is de afstand tussen twee punten een rechte lijn. In de praktijk hebben we meestal meer dan twee punten en/of is het pad tussen twee punten niet altijd een rechte lijn. Dat maakt de formule om de afstand te berekenen iets ingewikkelder maar de gedachte erachter blijft hetzelfde. Er zijn verschillende varianten maar daar gaan we hier niet nader op in. KNN wordt onder andere gebruikt voor inbraakdetectie. Hoewel het kan worden gebruikt voor regressieproblemen, wordt het meestal gebruikt voor classificatieproblemen.

De gedachte achter KNN is dat de waarden van de exogene variabelen in de testset worden vergeleken met de waarden van de exogene variabelen in de trainingset. De k waarnemingen waarvoor de overeenkomsten het grootst zijn (dus de afstand het kleinst) worden gebruikt om de waarde van de endogene variabele in de testset te voorspellen, bijvoorbeeld door over de k geselecteerde waarnemingen de gemiddelde waarde van de endogene variabele te berekenen (regressie) of de meest voorkomende waarde te kiezen (classificatie). De parameter k wordt vooraf gekozen.

Het KNN-algoritme is eenvoudig te implementeren, maar biedt weinig inzicht. Het is mogelijk om domeinkennis en producten van voorgaande ketenpartners als exogene variabelen in de analyse mee te nemen. Ook kan het resultaat weer als input dienen voor de prognoses van de volgende ketenpartner. Daarmee is dit algoritme ketenconsistent. Het nadeel is dat KNN niet goed kan omgaan met ruis in de data en daarmee gevoelig is voor uitschieters en fouten in de data. KNN vereist dus een hoge datakwaliteit, waardoor het wellicht minder geschikt is voor het PMJ.

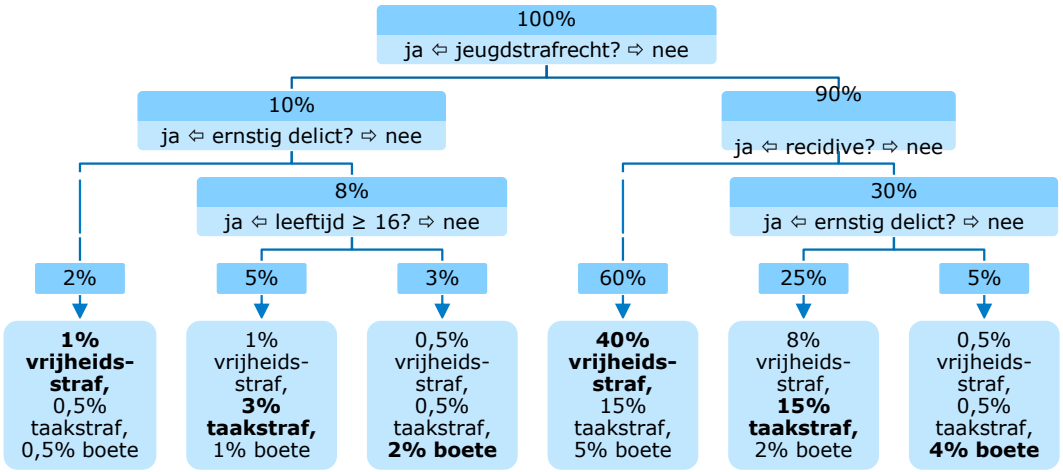
4.6.2 Beslisbomen

Een beslisboom ('decision tree') is een algoritme dat een boomstructuur bouwt. Beslissingsbomen worden over het algemeen ondersteboven gebouwd, zodat de bladeren zich onderaan de boom bevinden. Elk intern knooppunt ('internal node') is een beslissing of splitsing ten aanzien van een exogene variabele, elke tak ('branche') is het resultaat van die beslissing en elk blad ('leaf' of 'end node' of 'region') de consequentie van de beslissing voor de endogene variabele. In het geval van een classificatieboom is dit de meest voorkomende categorie of de kans op een bepaalde waarde van de endogene variabele en in het geval van een regressieboom is dit het gemiddelde over de endogene variabele. In box 4.4 staat een fictief voorbeeld.

Box 4.4 Fictief voorbeeld van een beslisboom voor gerechtelijke beslissingen

Figuur 4.4 laat een fictieve beslisboom zien. Op basis van een aantal kenmerken, zoals ernst van het delict, leeftijd en type strafrecht deelt het algoritme de instroom bij de rechtbank uiteindelijk in in zes significant van elkaar verschillende categorieën. De optimale wijze van splitsing wordt door het algoritme bepaald, bijvoorbeeld wat ernstige en wat niet ernstige delicten zijn en welke leeftijdsgrens precies moet worden gehanteerd. De kenmerken waarop het eerst wordt gesplitst zijn het belangrijkst. Naarmate men dieper in de boom komt, neemt het belang van het kenmerk af. Voor elk blad wordt vervolgens bepaald wat de meest voorkomende straf binnen dit blad is. Dat wordt dan de voorspelling voor zaken met genoemde kenmerken.

Figuur 4.4 Een fictieve beslisboom voor gerechtelijke beslissingen



Een beslisboom wordt geconstrueerd door de trainingset recursief te splitsen in subsets op basis van de waarden van de exogene variabelen totdat aan een stopcriterium is voldaan, zoals bijvoorbeeld de maximale diepte van de boom, minimale aantal waarnemingen van een blad, maximaal aantal knooppunten of wanneer de endogene variabele in een subset voor alle records dezelfde waarden heeft, of wanneer verder splitsen geen toegevoegde waarde heeft voor de voorspelkracht. Het algoritme bepaalt welke exogene variabelen moeten worden gebruikt om te splitsen, in welke volgorde en wat de drempelwaarden zijn. De bovenste knooppunten van de beslissingsboom zijn doorgaans het belangrijkst. Beslisbomen kunnen ondiep (‘shallow’) of diep (‘deep’) zijn. Ondiepe bomen hebben minder variantie maar een hogere bias. Diepe bomen hebben een lage bias maar een hoge variantie.

Beslisbomen hebben een aantal voordelen:

- de constructie ervan vereist geen domeinkennis;
- impliciete aannames worden vermeden;

- beslisbomen kunnen omgaan met een groot aantal exogenen en grote databestanden;
- beslisbomen zijn eenvoudig te begrijpen en uit te leggen.

Nadelen van beslisbomen zijn:

- Net als bij logistische regressie is de tijdscomponent, die nodig is om een groot aantal jaren vooruit te voorspellen, lastig te implementeren.
- Beslissingsbomen hebben de neiging hebben om te overfitten, dat wil zeggen dat het risico groot is dat elke unieke combinatie van de waarden van de exogene variabelen een eigen tak en blad heeft.
- Deze overfitting resulteert in een hoge variantie zodat een kleine variantie in de gegevens kan leiden tot een grote variantie in de resulterende prognose, waardoor een onstabiele boom ontstaat.
- Hierdoor zijn de prognoses van een endogene variabele met niet eerder waargenomen waarden van exogene variabelen, doorgaans vrij slecht. In de praktijk presteren veel andere methoden beter met vergelijkbare gegevens. Dit probleem kan worden verholpen door een enkele beslisboom te vervangen door een 'random forest' (zie paragraaf 6.2), waarbij meerdere beslisbomen worden geconstrueerd, maar dit is niet zo eenvoudig te interpreteren als één enkele beslisboom.

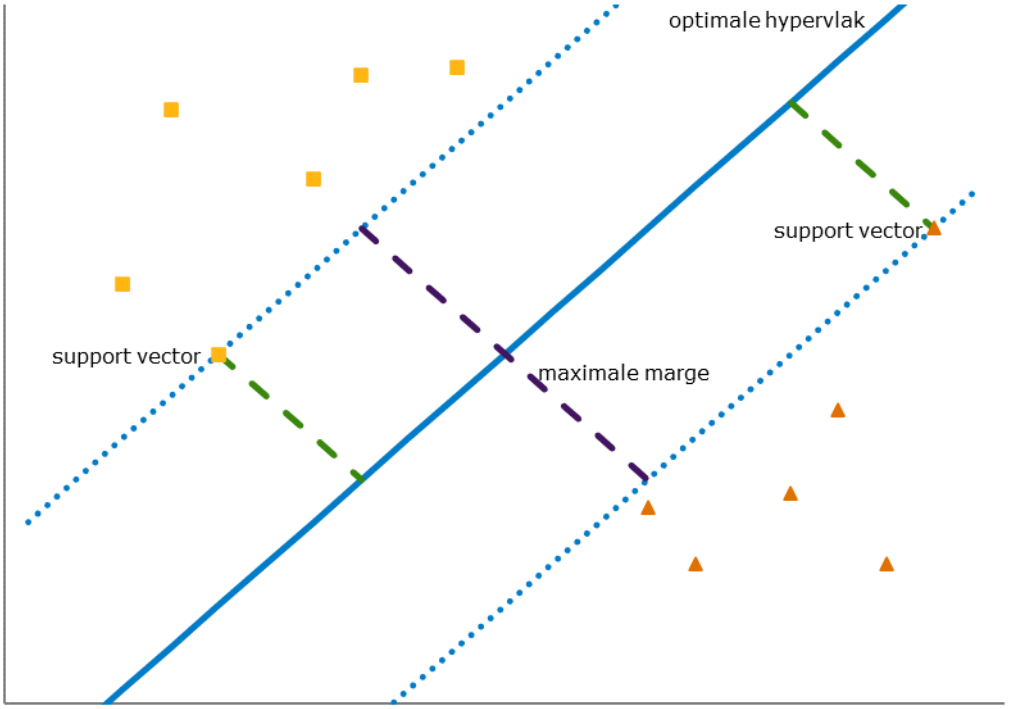
Een oplossing voor overfitting is het snoeien ('pruning') van een beslisboom. Nadat de beslisboom eerst volledig is opgebouwd, wordt elke tak aan de boom afzonderlijk onderzocht. Takken waarvan op basis van een criterium wordt aangenomen dat ze het gevolg zijn van overfitting, worden weggesnoeid. Een voorbeeld van een dergelijk criterium is het percentage foutieve prognoses in een validatieset.

4.6.3 Support vector machine/regressie

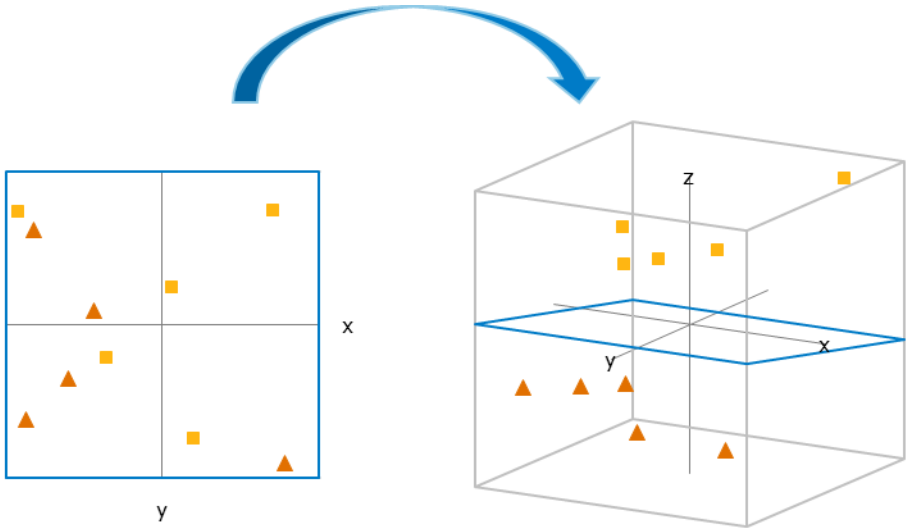
Support Vector Machine (SVM) is een lineair algoritme dat wordt gebruikt voor classificatie. Het doel van het SVM-algoritme is om een optimale combinatie van continue exogene variabelen te vinden waarvoor de categorieën van een binaire endogene variabele zo ver mogelijk van elkaar gescheiden worden. De ruimte tussen deze categorieën noemen we de maximale marge en het midden ervan het optimale hypervlak ('maximum-margin hyperplane'). Datapunten die op de grens van de marge liggen worden de 'support vectors' genoemd (zie figuur 4.5). Het doel van SVM is om een hypervlak te vinden die de marge maximaliseert. Als er twee exogene variabelen zijn, dan is dat een lijn. Als er drie exogene variabelen zijn, dan is dat een vlak. Bij meer dan drie exogene variabelen spreken we van een hypervlak.

Bij niet-lineaire problemen kan vaak in eerste instantie geen duidelijke scheiding worden aangebracht. In dat geval kan SVM een nieuwe exogene variabele creëren die een functie is van de oorspronkelijke exogene variabelen dusdanig dat niet-scheidbare problemen in scheidbare problemen worden omgezet (zie figuur 4.6). Dit wordt de 'kernel trick' genoemd. De functie heet de 'kernel' functie.

Figuur 4.5 Support vector machine



Figuur 4.6 Creatie van een extra exogene variabele in een support vector machine



Het voordeel van SVM is dat het veel exogene variabelen aan kan. SVM is veelzijdig omdat verschillende kernelfuncties kunnen worden gespecificeerd. SVM is robuust voor uitschieters, omdat SVM gebruikmaakt van een zogenoemde 'slack-variabele' die uitschieters of een zekere mate van overlap tussen de categorieën toelaat. Maar SVM is minder geschikt voor grotere datasets omdat de trainingstijd hoog kan zijn. SVM is ook minder effectief op datasets met veel ruis. Als er een kernelfunctie is gebruikt, zijn de prognoses gemaakt met SVM soms moeilijk te interpreteren en uit te leggen (afhankelijk van de complexiteit van de kernelfunctie). Hoewel SVM is ontworpen voor een endogene variabele met twee categorieën, bestaan er varianten voor meerdere categorieën of voor continue endogene variabelen (support vector regression, SVR).

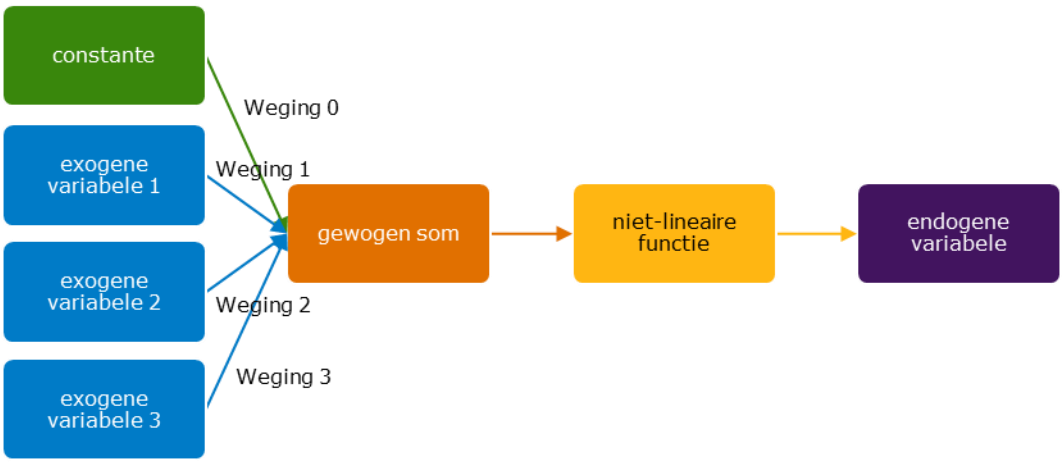
Op het eerste gezicht lijkt SVM op LDA (zie paragraaf 4.5.1). Beide algoritmes proberen om de afstand tussen categorieën van de endogene variabele zo groot mogelijk te maken op basis van de exogene variabelen. Toch zijn het heel verschillende algoritmes. SVM is discriminatief terwijl LDA generatief is (zie de inleiding van paragraaf 5.2). LDA veronderstelt dat alle gegevens normaal verdeeld zijn en dat alle categorieën identieke (co)variantie hebben. SVM doet helemaal geen aannames over de gegevens, wat het algoritme zeer flexibel maakt. Tegelijkertijd zorgt de flexibiliteit ervoor dat het vaak moeilijker is om de resultaten van een SVM-classificator te interpreteren in vergelijking met LDA. SVM kan gebruikmaken van kernelfuncties, LDA kan dat niet. LDA gebruikt de volledige dataset om covariantiematrices te schatten en is daardoor enigszins gevoelig voor uitschieters. Omdat SVM gebruikmaakt van een 'slack-variabele', is SVM ongevoeliger voor uitschieters.

4.6.4 Neurale netwerken

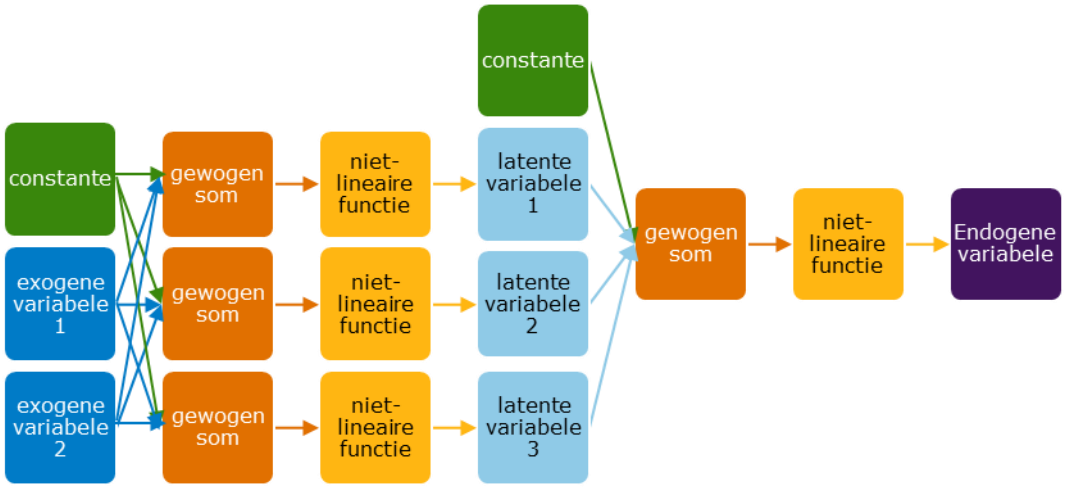
Neurale netwerken zijn geïnspireerd door hoe neuronen in de hersenen interacteren. Kunstmatige neurale netwerken (‘artificial neural networks’, ANN) worden gebruikt om patronen in complexe gegevens te vinden en zo gegevenspunten te voorspellen en te classificeren. Een neuraal netwerk kan uit verschillende lagen bestaan. Er is een inputlaag, eventueel één of meer verborgen lagen en een outputlaag. Elke laag bevat een aantal knooppunten (‘nodes’) die onderling met elkaar zijn verbonden. Deze verbindingen heten ‘edges’.

De meest simpele vorm van een neuraal netwerk is een ‘perceptron’ (zie figuur 4.7). Een perceptron kan beschouwd worden als een neuraal netwerk zonder verborgen lagen. In een perceptron wordt de gewogen som van de exogene variabelen plus een constante (ook wel bias genoemd) berekend en deze som dient vervolgens als input voor een zogenoemde activatiefunctie. De activatiefunctie kan een heel simpele of een heel complexe vorm aannemen maar de vorm wordt in belangrijke mate bepaald door het gewenste type uitkomst: een categorale variabele of een continue variabele. Het eindresultaat is een prognose van de endogene variabele. In een neuraal netwerk met één of meerdere verborgen lagen, zoals figuur 4.8, wordt elk knooppunt in de verborgen laag en de outputlaag berekend als een perceptron.

Figuur 4.7 Perceptron

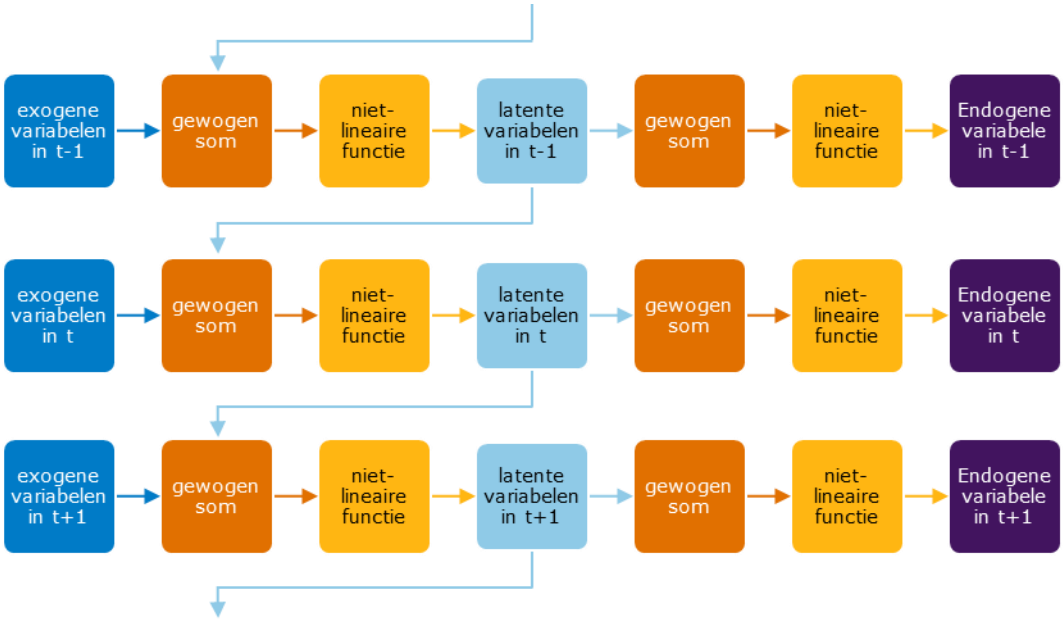


Figuur 4.8 Neuraal netwerk met 1 verborgen laag



Neurale netwerken worden getraind door record voor record de exogene variabelen in de trainingset aan het netwerk te voeden en een schatting van de endogene variabele te maken. Na elk record kan de schatting worden vergeleken met de werkelijke waarde van de endogene variabele. Vervolgens wordt op basis van dit verschil de wegingen van de exogene variabelen bijgesteld en wordt het volgende record aan het netwerk gevoed, totdat alle records in de trainingset zijn doorlopen. Er is ook een tijdreeksvariant, namelijk 'recurrent neural network' (RNN). Naast de reguliere exogene variabele wordt in een RNN ook de verborgen laag in de vorige periode als input doorgegeven aan de verborgen laag in de huidige periode (zie figuur 4.9).

Figuur 4.9 Recurrent neuraal netwerk met 1 verborgen laag



Het voordeel van een neuraal netwerk is dat alle exogene variabelen worden meegenomen in het model. Het algoritme bepaalt welke weging de exogene variabelen krijgen en er wordt geen informatie weggegooid. Door de verschillende lagen en de soms ingewikkelde activatiefuncties is de complexiteit erg hoog. Het eindresultaat is vaak niet goed herleidbaar naar specifieke exogene variabelen, want alles hangt met alles samen. Dit zorgt er ook voor dat het doorrekenen van een neuraal netwerk veel tijd kost. Wel kan de tijdscomponent worden ingebracht en is het mogelijk om een ketenconsistent netwerk te bouwen.

4.6.5 Samenvatting en implicaties voor het PMJ

Alle beschreven algoritmes in deze paragraaf zijn non-parametrisch, wat als voordeel heeft dat er vooraf weinig aannamen over de data worden gedaan. Verder verschillen de voor- en nadelen per algoritme. Met uitzondering van beslisbomen is het eindresultaat van de andere algoritmes vaak niet goed herleidbaar naar specifieke exogene variabelen. Het KNN-algoritme en beslisbomen zijn intuïtief en eenvoudig te berekenen. KNN is gevoelig voor uitschieters en fouten in de data. SVM is wel robuust voor uitschieters, maar ook minder effectief op datasets met veel ruis. Bij beslisbomen en SVM is het moeilijker om de tijdscomponent mee te nemen. Beslissingsbomen hebben de neiging tot overfitten. De trainingstijd van SVM en neurale netwerken is hoog. Alles overziend zouden KNN en beslisbomen een goede aanvulling voor het PMJ kunnen zijn, waarbij respectievelijk het nadeel van het niet kunnen herleiden naar specifieke exogene variabelen dan wel het ontbreken van de tijdscomponent moet worden afgewogen tegen de voordelen. De eigenschappen van de diverse algoritmes zijn in tabel 4.4 nog eens op een rijtje gezet.

Tabel 4.4 Kenmerken van algoritmes voor zowel regressie als classificatie

	K-nearest neighbours	Beslisbomen	Support vector machines/regression	Neurale netwerken
Algoritme	classificatie en regressie	classificatie en regressie	classificatie en regressie	classificatie en regressie
Endogene	Continu of categoraal	Continu of categoraal	Continu of categoraal	Continu of categoraal
Exogenen	Continu of categoraal	Continu of categoraal	Continu of categoraal	Continu of categoraal
Parametrisch	×	×	×	×/✓
Toepassingsgebied				
Inhoudelijke uitlegbaarheid van het algoritme	★★☆	★★★	★★☆	★☆☆
Eenvoud van het algoritme	★★★	★★☆	★★☆	★★☆
Implementeerbaarheid	★★★	★★★	★★★	★★☆
Rekentijd	★★★	★★★	★☆☆	★☆☆
Domeinkennis	✓	✓	✓	✓
Ketenconsistentie	✓	×	×	×
Tijdscomponent mogelijk	✓	×	×	✓
Inhoudelijke uitlegbaarheid van de prognoses	×	✓	×	×
Ruis in de data mogelijk	×	×	✓	×
Privacy: toepasbaar op geaggregeerde data	✓	✓	?	✓
indien microdata: aggregeerbaarheid	★★★ (continu) ★★☆ (categoraal)	★★★ (continu) ★★☆ (categoraal)	★★★ (continu) ★★☆ (categoraal)	★★★ (continu) ★★☆ (categoraal)
Rechtvaardigheid issues		neiging tot overfitten kan leiden tot ongewenste categorieën	kan leiden tot ongewenste categorieën	
Overig		Lage bias, hoge variantie; slechte prognoses van niet eerder waargenomen waarden		

★☆☆ moeilijk/slecht
★★☆ matig/gemiddeld
★★★ makkelijk/goed

5 Benutting van de steekproef

De benutting van de steekproef in het huidige PMJ is redelijk rechttoe rechtaan. Het PMJ wordt getraind op basis van jaargegevens. Voor sommige onderdelen zijn gegevens vanaf de jaren ’50 van de vorige eeuw beschikbaar, maar vanwege gewijzigde definities, beleid, wetgeving en omstandigheden worden de oudere gegevens minder relevant geacht voor de toekomst en daarom niet gebruikt voor de training van het model. Voor de meeste onderdelen beginnen de reeksen in de jaren ’90 van de vorige eeuw, maar het opsporingsgedeelte begint in de jaren ’70. Het model wordt getraind op basis van één trainingset. Er is een grote selectie aan exogene variabelen beschikbaar, maar in de praktijk komen slechts een beperkt aantal exogene variabelen in het empirisch model terecht. Deze worden geselecteerd op basis van modelselectiecriteria gebaseerd op de waarde van de aannemelijkheidsfunctie en het aantal parameters. Er vindt geen transformatie op de variabelen plaats anders dan de conversie naar (eerste verschillen van) logaritmes.
Het geringe aantal geselecteerde exogene variabelen kan het gevolg zijn van het beperkt aantal waarnemingen (bijvoorbeeld: vanaf 1990 t/m 2022 zijn dit er slechts 33) en/of de grote onderlinge afhankelijkheid van de exogene variabelen (bijvoorbeeld opleidingsniveau, werkloosheid, inkomen en economische groei zijn alle sterk aan elkaar gerelateerd). Door op een andere wijze met de data om te gaan kunnen deze problemen worden aangepakt, en kan de huidige steekproef wellicht meer benut worden. Er zijn twee populaire mogelijkheden die in combinatie met elkaar kunnen worden toegepast: dimensiereductie (paragraaf 5.1) en/of meerdere trainingsets te genereren uit één steekproef (paragraaf 5.2).

5.1 Dimensiereductie

Naarmate het aantal beschikbare exogene variabelen toeneemt, neemt het aantal dimensies van de gegevensverzameling toe. Naarmate het aantal dimensies toeneemt, neemt het aantal mogelijke combinaties van de exogene variabelen exponentieel toe en daarmee de complexiteit. Hierdoor neemt ook de hoeveelheid gegevens die nodig is om een goed resultaat te verkrijgen exponentieel toe. Dit kan weer leiden tot problemen zoals overfitting, langere rekentijd en verminderde nauwkeurigheid. Bovendien zijn sommige algoritmen gevoeliger voor het aantal dimensies dan andere. Dit staat bekend als de ‘curse of dimensionality’. Om de curse of dimensionality aan te pakken zijn er grofweg twee opties, namelijk een selectie van exogene variabelen maken of combinaties van exogene variabelen gebruiken. Paragraaf 5.1.1 beschrijft een aantal selectiemethodes terwijl paragraaf 5.1.2 de combinatiemethodes beschrijft.

5.1.1 Selectie van exogenen

De optimale keuze van de exogene variabelen is de selectie die resulteert in de kleinste validatiefout. Het is mogelijk om handmatig vooraf een selectie te maken. Het maken van een selectie vooraf heeft een aantal voordelen:

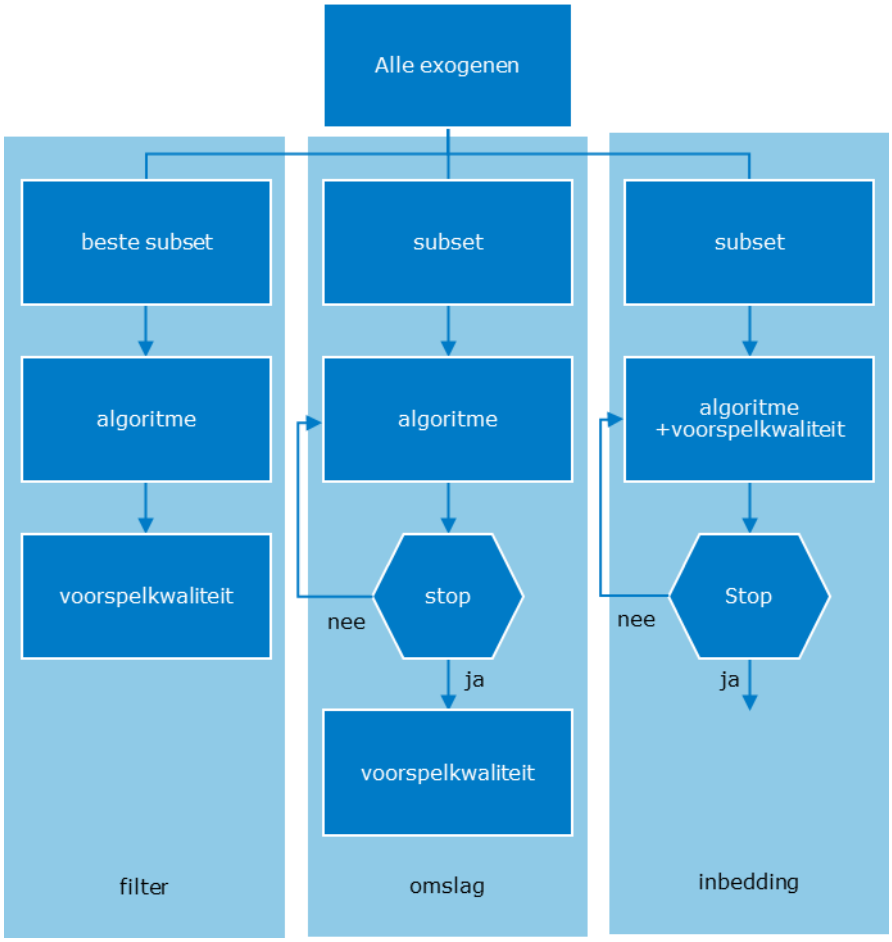
- Exogene variabelen die waarschijnlijk irrelevant zijn voor het te onderzoeken probleem (bijv. op basis van theorie en/of empirie), kunnen worden weggelaten. Hierdoor is er minder ruis in het model, hetgeen de kwaliteit van de prognoses ten goede komt.

- Het verwijderen van gecorreleerde exogene variabelen zal de interpreteerbaarheid van het model verbeteren.
- Een kleinere dataset betekent dat het algoritme sneller wordt getraind.

Maar het is ook mogelijk om de selectie te automatiseren, wat als voordeel heeft dat naar meer of bredere selecties gekeken kan worden. Er zijn drie verschillende technieken voor selectie van de exogene variabelen: filtermethoden, omslagmethoden ('wrapper') en ingebedde methoden (zie figuur 5.1). Filtermethoden gebruiken geen specifiek algoritme, maar kijken naar de correlatie tussen de exogene variabelen en de endogene variabele. Hiervoor zijn in de loop der jaren vele criteria ontwikkeld. De keuze hiertussen is afhankelijk van het type endogene en exogenen variabelen (categoraal of continu). Bij de omslagmethoden wordt een specifiek algoritme toegepast op verschillende selecties van de exogene variabelen. Op basis van een vooraf gekozen criterium worden de prognoses met elkaar vergeleken. De selectie van exogene variabelen die de beste prognose genereert is de optimale selectie. Een veel gebruikte omslagmethode is stapsgewijze regressie ('step wise regression') of classificatie en deze methode wordt hieronder nader toegelicht. Maar ook modelselectiecriteria gebaseerd op de waarde van de aannemelijkheidsfunctie en het aantal parameters kunnen als omslagmethode worden beschouwd. Bij ingebedde methoden is de selectie van exogene variabelen een integraal onderdeel van het algoritme, zoals bijvoorbeeld regularisatiemethoden zoals LASSO en elastic net (zie paragraaf 3.1.4) en beslisbomen (zie paragraaf 4.6.2).

Stapsgewijze regressie
Het doel van stapsgewijze regressie is om de optimale subset van exogene variabelen te vinden. Een optie is om het theoretische model toe te passen op alle mogelijke combinaties van exogene variabelen. Vervolgens wordt het model gekozen dat voldoet aan de vooraf gekozen criteria (bv. een modelselectiecriterium gebaseerd op de waarde van de aannemelijkheidsfunctie en het aantal parameters of een voorspelcriterium als gemiddelde kwadratische fout). Het probleem met deze methode is dat het aantal mogelijke combinaties exponentieel toeneemt met het aantal exogene variabelen. Een alternatief is om op iteratieve wijze een subset van exogene variabelen te vinden. Hiervoor zijn twee mogelijkheden. Bij voorwaartse stapsgewijze selectie van exogene variabelen wordt begonnen met een kleine subset van de exogene variabelen. Daarna wordt het model opnieuw getraind met één extra exogene variabele. Dit gebeurt voor alle resterende exogene variabelen. Vervolgens wordt het model gekozen dat het best voldoet aan de vooraf gekozen criteria. Met dit model wordt de hierboven beschreven herhaald totdat aan een bepaald convergentiecriterium is voldaan. Bij achterwaartse stapsgewijze selectie werkt het proces precies andersom. Er wordt begonnen met een model dat alle exogene variabelen bevat en één voor één wordt er een variabele verwijderd. Merk op dat voorwaartse en achterwaartse stapsgewijze selectie van variabelen niet noodzakelijkerwijs resulteren in de optimale subset van exogene variabelen.

Figuur 5.1 Filter-, omslag- en ingebedde methoden



5.1.2 Combinatie van exogenen

In plaats van een selectie van exogene variabelen mee te nemen in het algoritme, kunnen we meerdere exogene variabelen combineren tot één variabele. De gecombineerde variabelen kunnen vervolgens worden gebruikt als input voor een regressie of classificatie-algoritme zoals besproken in hoofdstuk 3 en 4. Omdat het aantal combinaties doorgaans lager is dan het oorspronkelijk aantal exogene variabelen wordt het aantal dimensies gereduceerd, terwijl de informatie die verloren gaat minimaal is. Een ander voordeel is dat een deel van de ruis hiermee uit de data wordt gehaald waardoor het model minder zal overfitten. De reductie in het aantal dimensies leidt tevens tot een versnelling van het trainingsproces van het model. Het nadeel van het gebruik van dergelijke combinaties in een algoritme is dat de resulterende prognose van de endogene variabele moeilijker te interpreteren is, want de bijdrage van specifieke exogene variabelen aan de prognose van de endogene variabele kan lastig zijn om te achterhalen. Er zijn meerdere methoden, maar in dit hoofdstuk bespreken we een aantal populaire methoden, namelijk principale componentenanalyse en factor analyse. Ook de eerder besproken LDA, QDA, FDA (zie paragraaf 4.5.1) kunnen worden gebruikt om exogene variabelen te combineren en daarmee het aantal dimensies te reduceren.

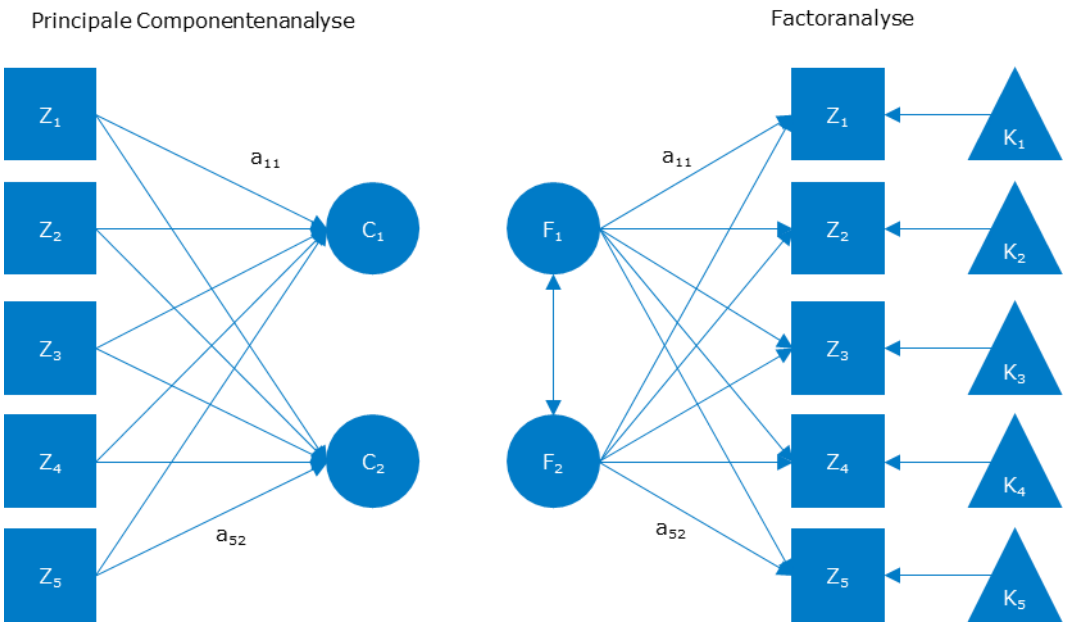
Principale Componenten Analyse
Het belangrijkste doel van Principale Componenten Analyse (PCA) is om het aantal dimensies (d.w.z. het aantal exogene variabelen) van een gegevensverzameling te verminderen en tegelijkertijd de belangrijkste patronen of relaties tussen de exogene variabelen te behouden zonder enige voorafgaande kennis van de endogene variabele. PCA maakt lineaire combinaties van de gestandaardiseerde (zie box 3.2) exogene variabelen die loodrecht op elkaar staan. Het totaal aantal lineaire combinaties is gelijk aan het aantal exogene variabelen. De totale variantie van alle principale componenten is gelijk aan de totale variantie van alle originele exogene variabelen. De lineaire combinaties zijn gerangschikt naar de mate van verklaarde variantie tussen de exogene variabelen. De eerste combinatie verklaart de meeste onderlinge variantie tussen de exogene variabelen, de laatste de minste. Principale componenten die weinig variantie verklaren kunnen worden weggelaten, waardoor het aantal dimensies wordt gereduceerd. Soms kunnen de afzonderlijke componenten worden geassocieerd met bepaalde maatschappelijk ontwikkelingen. Bijvoorbeeld, als de economische variabelen, zoals werkloosheid en bbp, sterk doorwegen in component 1, dan kan component 1 worden beschouwd als een indicatie van economische ontwikkeling. Dit is echter lang niet altijd evident. Verder kan een component die weinig variantie tussen de exogene variabelen verklaart, wel degelijk een hoge correlatie vertonen met een endogene variabele, bijvoorbeeld als de component een heel specifiek fenomeen meet, die alleen maar van belang is voor één specifieke endogene variabele. In dat geval is de reductie van het aantal dimensies beperkt.

Als de exogene variabelen binair zijn (d.w.z dat ze de waarden 0 of 1 kunnen aannemen), dan is PCA minder geschikt. Het alternatief is correspondentieanalyse (CA; ook wel reciproke middeling genoemd). CA is conceptueel vergelijkbaar met PCA, maar is van toepassing op categorale gegevens in plaats van op continue gegevens. CA vergelijkt twee binaire exogene variabelen. Als meer dan twee variabelen moeten worden vergeleken, dan moet worden gekozen voor meervoudige correspondentie-analyse (MCA).

Factoranalyse
Factoranalyse wordt ook gebruikt om het aantal dimensies te reduceren maar werkt fundamenteel anders dan PCA. Het doel van factoranalyse is het blootleggen van de latente fenomenen (ook wel factoren genoemd) achter de exogene variabelen. De gedachte erachter is dat een deel van de gestandaardiseerde (zie box 3.2) exogene variabelen een gezamenlijke, niet-waargenomen oorzaak heeft. Ook kunnen er latente fenomenen zijn die niet meetbaar zijn met één enkele exogene variabele. Factoranalyse probeert deze achterliggende oorzaken te kwantificeren. De totale variantie van alle factoren is niet gelijk aan de totale variantie van alle originele exogene variabelen. Elke exogene variabele kent een uniek deel. Het aantal factoren is altijd kleiner dan het aantal exogene variabelen. Een voorbeeld van een factor is de sociaal-economische groep waartoe iemand behoort. Exogene variabelen die daarmee geassocieerd kunnen worden, zijn bijvoorbeeld opleidingsniveau en werkloosheid, die ook onderling gecorreleerd zijn.

Er zijn veel verschillen tussen factor analyse en principale componenten analyse. Maar het belangrijkste verschil is dat principale componenten in PCA worden berekend als de lineaire combinaties van de oorspronkelijke exogene variabelen, terwijl in factoranalyse de oorspronkelijke exogene variabelen worden gedefinieerd als lineaire combinaties van de factoren (zie figuur 5.2).

Figuur 5.2 Principale componentenanalyse versus factoranalyse



Z=gestandaardiseerde exogene variabele, C=principale component, F=gemeenschappelijke factor, K=unieke factor, a=component- of factorlading.

5.2 Genereren van meerdere trainingsets

Om goed te functioneren hebben de algoritmes uit hoofdstuk 4 en 5 veel data nodig. Helaas is er soms niet zo veel data beschikbaar. Daarom zijn er technieken ontwikkeld om uit één steekproef meerdere trainingsets te genereren (‘resampling’) De twee populairste methoden worden hier besproken. Paragraaf 5.2.1 gaat in op kruisvalidatie en paragraaf 5.2.2 gaat in op bootstrapping.

5.2.1 Kruisvalidatie

Het primaire doel van kruisvalidatie is het valideren van het empirisch model met behulp van een maatstaf die niet op slechts één willekeurige validatieset is gebaseerd maar op meerdere validatiesets. Stel dat van de totale gegevensverzameling eerst een testset wordt afgesplitst. De resterende waarnemingen worden vervolgens opgedeeld in een trainingset en een validatieset (zie paragraaf 4.1.2). Bij een willekeurige splitsing is de vraag altijd hoe representatief deze sets zijn. Het kan zijn dat door puur toeval de trainingset uitschieters bevat, die niet in de validatieset zitten of andersom. Het kan ook voorkomen dat niet alle categorieën van de endogene of exogene variabelen in de trainingset en/of de validatieset voorkomen, maar wel in de testset. Bovendien kan bij een klein aantal waarnemingen de omvang van de trainingset zeer klein worden, wat kan resulteren in een slecht model of uitsluiting van bepaalde algoritmes. Een oplossing voor dit probleem is kruisvalidatie (‘cross validation’). Bij kruisvalidatie wordt de gegevensverzameling niet slechts één keer in een trainingset en een validatieset gesplitst maar herhaaldelijk gesplitst. Hier worden twee varianten besproken, namelijk ‘Leave-one-out cross-validation’ (LOOCV) en ‘k-fold cross-validation’ (KFCV).

Bij de LOOCV methode wordt de validatieset gevuld met de eerste waarneming, terwijl alle andere waarnemingen naar de trainingset gaan. Daarna wordt de endogene variabele met een of ander algoritme geschat en de maatstaf die wordt gebruikt om het model te valideren opgeslagen. Dit is meestal de gemiddelde kwadratische fout. Vervolgens wordt de tweede waarneming in de validatieset geplaatst en de rest in de trainingset en wordt het model opnieuw geschat en de voorspelmaatstaf opgeslagen. Dit proces wordt herhaald totdat alle waarnemingen een keer in de validatieset hebben gezeten. Daarna kan op basis van de opgeslagen resultaten de gemiddelde validatiefout worden berekend.

Bij de KFCV-methode wordt de gegevensverzameling (na afsplitsing van de testset) verdeeld in k groepen van gelijke omvang. Vervolgens wordt de validatieset gevuld met de eerste groep waarnemingen, terwijl alle andere waarnemingen naar de trainingset gaan. Daarna wordt de endogene variabele met een of ander algoritme geschat en de gekozen voorspelmaatstaf opgeslagen. Vervolgens wordt dit proces herhaald door de volgende groep waarnemingen in de validatieset te plaatsen en de rest in de trainingset totdat alle k groepen van waarnemingen een keer in de validatieset hebben gezeten. Tot slot kan op basis van de opgeslagen resultaten de gemiddelde validatiefout worden berekend. Merk op dat de LOOCV-methode een speciaal geval is van de KFCV-methode, namelijk wanneer iedere groep één waarneming bevatten.

In vergelijking met een willekeurige splitsing in een trainingset en validatieset is het voordeel van de LOOCV-methode dat de validatiefout niet wordt overschat en dat er geen willekeur is in de splitsing. Een nadeel van deze methode is dat het tijdrovender is. Een ander nadeel is dat de resultaten sterk gecorreleerd zijn omdat ze met vrijwel dezelfde steekproeven zijn geschat, waardoor de variantie vrij hoog is. De KFCV-methode zit qua rekentijd, validatiefout, correlatie en variantie tussen de LOOCV-methode en de willekeurige splitsing in. Kruisvalidatie op tijdreeksen leidt tot gaten in de tijdreeks wat tot problemen kan leiden bij het schatten van de autocorrelatie.

5.2.2 Bootstrapping

Een andere manier om meerdere trainingsets te genereren is ‘bootstrapping’. Het primaire doel van bootstrapping is om de empirische kansverdeling van de parameters van het model vast te stellen en daarmee een robuuste schatting te maken van de standaardfouten en de betrouwbaarheidsintervallen. Maar bootstrapping kan ook worden gebruikt voor het combineren van prognoses van hetzelfde algoritme (ofwel ‘bagging’, zie paragraaf 6.2).

Stel dat de gegevensverzameling gesplitst is in een trainingset, validatieset en testset. Bij bootstrapping wordt een willekeurige steekproef uit de trainingset getrokken met teruglegging. Teruglegging betekent dat een item nadat deze is getrokken weer teruggelegd wordt in de ‘pot’ zodat deze nog een keer getrokken kan worden. De getrokken steekproef is altijd even groot als de originele trainingset. Als de steekproef heel groot is, dan zal circa 63% van de items uit de trainingset in de steekproef zitten. Het is dus zeer waarschijnlijk dat de steekproef dubbele items bevat.

Bootstrapping kan op verschillende momenten in het proces worden toegepast. Als bootstrapping wordt toegepast op de oorspronkelijke waarnemingen, dan is er sprake van ‘case resampling’. Deze steekproef kan vervolgens worden gebruikt om de

endogene variabele met één of ander algoritme te schatten. Maar het is ook mogelijk om eerst op basis van de oorspronkelijke trainingset een model te schatten met één of ander algoritme en dan de residuen, dat wil zeggen het verschil tussen werkelijke en voorspelde waarden, te berekenen. Vervolgens wordt bootstrapping toegepast op de residuen en wordt een nieuwe endogene variabele gegenereerd die de som is van de schatting van de endogene variabele en het gebootstrapte residu. Tot slot kan dan hetzelfde algoritme weer worden toegepast op de nieuwe endogene variabele. Dit heet 'residual bootstrapping'. Voorwaarde is dat de residuen onafhankelijk van elkaar zijn en uit dezelfde verdeling komen.

De hierboven beschreven bootstrap methoden werken niet voor tijdreeksen omdat de waarnemingen in een tijdreeks in de tijd met elkaar gecorreleerd zijn. Een oplossing daarvoor is 'block bootstrapping'. Bij block bootstrapping worden de waarnemingen eerst in blokken van gelijke omvang opgedeeld. Vervolgens wordt een willekeurige steekproef van blokken (in plaats van waarnemingen) getrokken met teruglegging. Eventueel kunnen de blokken elkaar overlappen. In dat geval is er sprake van 'moving block bootstrapping'. Het voordeel van deze methoden is dat een deel van de temporele correlatie behouden blijft. Een eenvoudiger manier om de temporele correlatie te behouden is om vertraagde endogene variabelen als exogene variabelen te beschouwen en deze expliciet mee te nemen in gewone bootstrapping.

5.3 Samenvatting en implicaties voor het PMJ

Er zijn meerdere methoden om de huidige steekproef ruimer te benutten. Zo kunnen in plaats van individuele exogene variabelen combinaties van exogene variabelen in het model worden opgenomen, bijvoorbeeld door middel van PCA of factoranalyse. Het voordeel is dat er zo min mogelijk informatie verloren gaat terwijl tegelijkertijd het aantal input variabelen wordt gereduceerd. Maar er kleven ook een aantal nadelen aan. De resulterende prognoses zijn niet goed herleidbaar naar specifieke exogene variabelen en er is een reëel risico dat in het empirische model uiteindelijk alles met alles samenhangt. Ook is het mogelijk dat een exogene variabele die naar verwachting weinig invloed heeft op bepaalde justitiële onderdelen, toch enige mate van invloed uitoefent omdat deze variabele nu eenmaal onderdeel is van een combinatie. Bijvoorbeeld als de bevolking van 12 t/m 17 jaar (met een kleine lading) onderdeel is van de gecombineerde variabele Z en Z heeft een significante invloed op het aantal maatregelen voor plaatsing in een inrichting voor stelselmatige daders (ISD), dan krijgen we de vreemde situatie dat de ontwikkeling in het aantal 12- t/m 17-jarigen het aantal ISD-maatregelen beïnvloedt, terwijl deze bevolkingsgroep hiervoor helemaal niet aanmerking komt. Vanwege deze nadelen is de toepassing van PCA of factoranalyse niet wenselijk voor het PMJ.

Een andere manier om de steekproef breder te benutten is om de validatiefouten of de standaardfouten en de betrouwbaarheidsintervallen van de geschatte parameters te verbeteren. Hiertoe worden uit dezelfde steekproef meerdere trainingsets en validatiesets getrokken waarmee iedere keer hetzelfde model wordt geschat. Deze technieken hebben geen invloed op de parameters van het uiteindelijke empirische model zelf, want deze worden geschat over de volledige trainingset (zonder dubbelingen). Wel kunnen deze methoden gebruikt worden om verschillende theoretisch modellen en/of algoritmes met elkaar te vergelijken. Maar dan moet kruisvalidatie of bootstrapping op alle theoretische modellen en algoritmes worden toegepast. Bij een groot aantal modellen, algoritmes en/of waarnemingen kan dit zeer

rekenintensief zijn. Mede omdat het PMJ doorgaans niet wordt beoordeeld op één specifieke maatstaf maar op meerdere maatstaven en op theoretische juistheid lijkt het marginale voordeel van betere voorspelmaatstaven en kennis over de verdeling van de parameters niet op te wegen tegen het nadeel van de grotere rekenintensiteit.

6 Combineren van prognoses

Combinatietechnieken ('ensembling') zijn methoden die de prognoses van een specifieke endogene variabele uit verschillende voorspellende modellen, algoritmen en/of steekproeven combineren om daarmee de nauwkeurigheid van de prognose te verbeteren en de bias en/of variantie te verkleinen. In dit hoofdstuk worden een aantal veel voorkomende opties besproken.

6.1 Ensemble averaging

De meest voor de hand liggende methode is om met verschillende algoritmes dezelfde endogene variabele te voorspellen en het (gewogen) gemiddelde over deze prognoses te berekenen. Dit wordt 'ensemble averaging' genoemd (zie voor een overzicht van de afgelopen 50 jaar Wang et al., 2023). De eventuele weging kan bijvoorbeeld worden bepaald door te kijken hoe goed het gevonden model de waarden van de endogene variabele in de testset voorspelt. In het geval van classificatieproblemen wordt de categorie gekozen die het vaakst wordt voorspeld ('plurality voting'). Dit wordt 'hard voting' genoemd. Als de classificatie-algoritmes een kans voorspellen, kunnen de geschatte kansen voor elke categorie worden opgeteld, waarna de categorie met de grootste som wordt gekozen. Dit staat bekend als 'soft voting'. In de praktijk blijkt dat ensemble averaging vaak leidt tot een betere prognose in vergelijking met de prognose uit één enkel algoritme, omdat de diversiteit van de algoritmes waarschijnlijk resulteert in een kleinere variantie.

6.2 Bagging

De 'bagging'-methode (ook wel 'bootstrap aggregating' genoemd) combineert prognoses van hetzelfde algoritme die op verschillende splitsingen van de trainingsets zijn toegepast. Het doel is om de variantie te verkleinen. Bij bagging wordt hetzelfde algoritme toegepast op een groot aantal gebootstrapte steekproeven (zie paragraaf 5.2.2). Vervolgens wordt het gemiddelde van alle prognoses berekend. Omdat de focus ligt op de vermindering van de variantie is deze methode vooral geschikt voor algoritmes met lage bias en hoge variantie. Paragraaf 6.2.1. gaat in op een bijzonder geval van bagging, namelijk 'random forest' ofwel bagging van beslisbomen.

6.2.1 Random forest

Bij 'random forest' is bagging een essentieel onderdeel van het algoritme. Random forest is een woud van diepe beslisbomen (zie paragraaf 4.6.2), waarbij elke beslisboom op een andere gebootstrapte steekproef van waarnemingen wordt opgebouwd. Ook wordt bij iedere boom een willekeurige selectie van de exogene variabelen gemaakt, waarop vervolgens de splitsing wordt gemaakt. Dit voorkomt dat bepaalde exogene variabelen dominant zijn, vermindert de correlatie tussen de prognoses van de verschillende beslisbomen en maakt de beslisboom minder gevoelig voor missende gegevens. Als het regressiebomen zijn, dan wordt het gemiddelde van de prognoses berekend. Als het classificatiebomen betreft, dan wordt de categorie gekozen die het vaakst wordt voorspeld.

6.3 Boosting

'Boosting'-methoden werken ongeveer hetzelfde als bagging-methoden. Prognoses van hetzelfde algoritme worden gecombineerd. Maar in tegenstelling tot bagging is boosting gericht op het verminderen van de bias. Bij boosting wordt hetzelfde algoritme iteratief toegepast op een adaptieve manier: in elke iteratie wordt rekening gehouden met de waarnemingen die niet goed werden voorspeld in de vorige iteratie. Boosting kan, net als bagging, zowel voor regressie als voor classificatieproblemen worden gebruikt. Omdat de focus ligt op de vermindering van de bias is boosting vooral geschikt voor algoritmes met hoge bias en lage variantie. Er zijn verschillende vormen van boosting zoals 'adaptive boosting' en 'gradient boosting'. De verschillen zitten vooral in de wijze waarop de adaptieve aanpassing wordt gedaan. De adaptieve aanpassing vereist wel een hoge datakwaliteit. Uitschieters en ruis in de data zijn problematisch en kunnen ertoe leiden dat men in een virtuele kuil terecht komt waar men alleen met onrealistisch aanpassingen uit kan komen. Ook is het algoritme rekenintensief.

6.4 Stacking

Bij 'stacking'-methoden worden de prognoses van verschillende algoritmes gecombineerd in een metamodel dat vervolgens leidt tot een metaprognose. Daarvoor wordt de trainingset in tweeën gesplitst. Het eerste deel wordt gebruikt voor de verschillende algoritmes. De modellen die hiermee geschat worden, worden gebruikt om de endogene variabele in het tweede deel van de trainingset te voorspellen. Tot slot wordt er een metamodel geschat op het tweede deel van de trainingset, waarbij de voorspelde waarden input zijn (en dus niet de exogene variabelen). Een duidelijk nadeel van deze opsplitsing is dat slechts de helft van de trainingset beschikbaar is voor het trainen van de modellen. Dit probleem kan eventueel opgelost worden door kruisvalidatie toe te passen (zie paragraaf 5.2.1) Voor een classificatieprobleem kunnen bijvoorbeeld op het eerste deel van de trainingset een KNN-classificator, een logistische regressie en een SVM worden toegepast en op het tweede deel van de trainingset een neuraal netwerk of lineaire regressie als metamodel.

6.5 Samenvatting en implicaties voor het PMJ

Er zijn een aantal manieren waarop verschillende prognoses van dezelfde endogene variabele kunnen worden gecombineerd. Bij bagging en boosting worden prognoses van hetzelfde algoritme maar verschillende steekproeven gecombineerd. Met name bagging zou voor het PMJ interessant kunnen zijn, omdat het relatief eenvoudig te implementeren is en het model zelf ook niet al te complex wordt, zodat dat de resultaten nog steeds uitlegbaar zijn. Boosting is minder interessant als de econometrische benadering wordt gevolgd omdat deze benadering ervanuit gaat dat er een unbiased schatter wordt gekozen.

Bij ensemble averaging en stacking kunnen prognoses van modellen geschat met verschillende algoritmes worden gecombineerd. Met name ensemble averaging is een veelbelovende techniek voor het PMJ, omdat het relatief eenvoudig te implementeren is en er sowieso in de testfase vaak al verschillende algoritmes worden uitgeprobeerd. Tot nu toe werd op basis van verschillende criteria uiteindelijk één model gekozen, hoewel de onderlinge verschillen in voorspelkwaliteit vaak gering waren. Als alternatief zouden meerdere algoritmes kunnen worden gebruikt om dezelfde endogene variabele

te voorspellen en hierover het gemiddelde worden berekend. Zo zou bijvoorbeeld de instroom bij het OM kunnen worden voorspeld op basis van het aantal verdachten dat door de politie wordt geregistreerd, maar ook door middel van een tijdreeksanalyse. De uiteindelijke prognose wordt dan het gemiddelde van beide algoritmes. Stacking is ook een optie zolang het metamodel niet te ingewikkeld is. Maar het nadeel is dat het veel rekentijd kost en dat er (relatief) veel data nodig is

7 Conclusie en aanbevelingen

In hoofdstuk 3 tot en met 6 zijn een aantal algoritmes aan de orde gekomen die in potentie een nuttige bijdrage aan het PMJ zouden kunnen leveren. Welke van deze algoritmes daadwerkelijk een kans van slagen hebben, hangt mede af van de randvoorwaarden waaronder het PMJ moet worden gebouwd en geactualiseerd (paragraaf 7.1) en de criteria waarop de algoritmes moeten worden beoordeeld (paragraaf 7.2). De implicaties voor het PMJ staan in paragraaf 7.3. Paragraaf 7.4 doet concrete aanbevelingen voor de toekomst.

7.1 Randvoorwaarden

Eind jaren '90 waren de belangrijkste redenen voor de ontwikkeling van het PMJ (Van der Heijden, 2002):

- meer inzicht in de ramingen;
- meer samenhang en consistentie in de ramingen;
- meer uniformiteit en statistische kwaliteit van de ramingen.

Dit doel is grotendeels bereikt, wat is bevestigd in vele externe evaluaties[6] in de afgelopen 25 jaar. Desalniettemin bieden de ontwikkelingen op het terrein van machine learning en computertechnologie nieuwe mogelijkheden. In hoofdstukken 3 t/m 6 zijn een groot aantal technieken beschreven en beoordeeld op de criteria genoemd in hoofdstuk 1. Maar om bruikbaar te zijn voor het PMJ, moeten de technieken ook aan een aantal randvoorwaarden voldoen, die vooral voortkomen uit het begrotingsproces en de wensen van de eindgebruikers van de PMJ-ramingen:

- Het empirische model moet ketenconsistent zijn. De prognose van de uitstroom van de ene ketenpartner moet doorwerken in de prognose van de instroom van de daaropvolgende ketenpartner.
- Er moet zeven jaar vooruit voorspeld kunnen worden, d.w.z. de begrotingshorizon (vijf jaren) plus de twee jaren tussen het laatst bekende realisatiejaar en het eerste begrotingsjaar.
- De prognoses moeten inhoudelijk uitlegbaar zijn. Beleidsmakers willen graag kunnen begrijpen waarom de prognoses zijn zoals ze zijn. In de praktijk betekent dit dat ze herleidbaar moeten zijn naar concrete inputvariabelen en dat de geschatte relatie een zekere mate van logica of causaliteit moet bevatten.
- Vanwege de planning van het begrotingsproces moeten de parameters van het model jaarlijks in november geactualiseerd zijn. Omdat een aantal gegevens pas eind september beschikbaar is, betekent dit in de praktijk dat de actualisering binnen een periode van circa zes weken moet plaatsvinden.
- Het gekozen algoritme moet rechtvaardig zijn. Gemaakte keuzes of beslisregels mogen er niet onbedoeld toe leiden dat het algoritme een discriminerend karakter krijgt.

[6] Theeuwes & De Winter (1998), KPMG/BEA (1998), Spapens et al. (2001), Bomhoff et al. (2002), Biermans & Van Leeuwen (2003), Goudriaan (2004), Felsö et al. (2006), Bont et al. (2009), Everhardt et al. (2016), De Poot et al. (2020).

7.2 Beoordelingscriteria

De criteria waarop de kwaliteit van een empirisch model en de daaruit voortkomende prognoses worden beoordeeld zijn afhankelijk van de gekozen benadering: econometrie of machine learning. Hoewel er grote overlap is tussen de technieken die in de econometrie en in machine learning worden gebruikt, werkt de econometrie meer vanuit de theorie en is machine learning meer datagedreven. Dit leidt tot andere beoordelingscriteria.

7.2.1 De ene voorspelfout is de andere niet

Er zijn twee type voorspelfouten: 'false positives' en 'false negatives' (zie figuur 7.1). In het geval van classificatie zijn false positives voorspelfouten waarbij ten onrechte wordt voorspeld dat een bepaalde gebeurtenis zich voordoet, bijvoorbeeld een verdachte die schuldig wordt bevonden, terwijl hij dat niet is. False negatives zijn voorspelfouten waarbij ten onrechte wordt voorspeld dat een bepaalde gebeurtenis zich niet voordoet, bijvoorbeeld een verdachte die niet schuldig wordt bevonden, terwijl hij dat wel is. Idealiter worden beide type voorspelfouten geminimaliseerd, maar in de praktijk zijn false positives and false negatives niet helemaal te voorkomen. Welk type fout erger is, hangt af van de probleemstelling. Zo zal in het bovengenoemde voorbeeld de maatschappelijke schade van ten onrechte veroordeelde verdachten moeten worden afgewogen tegen de maatschappelijke schade van ten onrechte vrijgesproken verdachten. Dit bepaalt of de false positives of de false negatives of beide moeten worden geminimaliseerd.

Figuur 7.1 Verschillende type fouten*

		Geobserveerde waarden		
		Positive (P)	*Negative (N)*	
Voorspelde waarden	*positive (P)*	true positive (TP)	false positive (FP) (type I fout) loos alarm, overschatting	precision= aantal TP's / aantal voorspelde P's
	negative (N)	false negative (FN) (type II fout), gemiste signalen, onderschatting	true negative (TN)	
		recall= aantal TP's / aantal geobserveerde P's	false positive rate= aantal FP's / aantal geobserveerde N's	

* In het Engels wordt deze matrix een 'contingency table' of 'confusion matrix' genoemd. Merk op dat de termen 'positive' en 'negative' niets met het teken van een variabele of getal te maken hebben. Deze termen zijn afkomstig uit de gezondheidszorg waarbij een patiënt als ziek (positieve testuitslagen) of niet ziek (negatieve testuitslagen) wordt gediagnosticeerd.

In het geval van regressie kan zich een vergelijkbaar probleem voordoen. Te lage voorspellingen (false negatives) kunnen problematischer zijn dan te hoge voorspellingen (false positives) omdat organisaties doorgaans een onverwacht overschot makkelijker kunnen opvangen dan een onverwacht tekort. Ook in dit geval

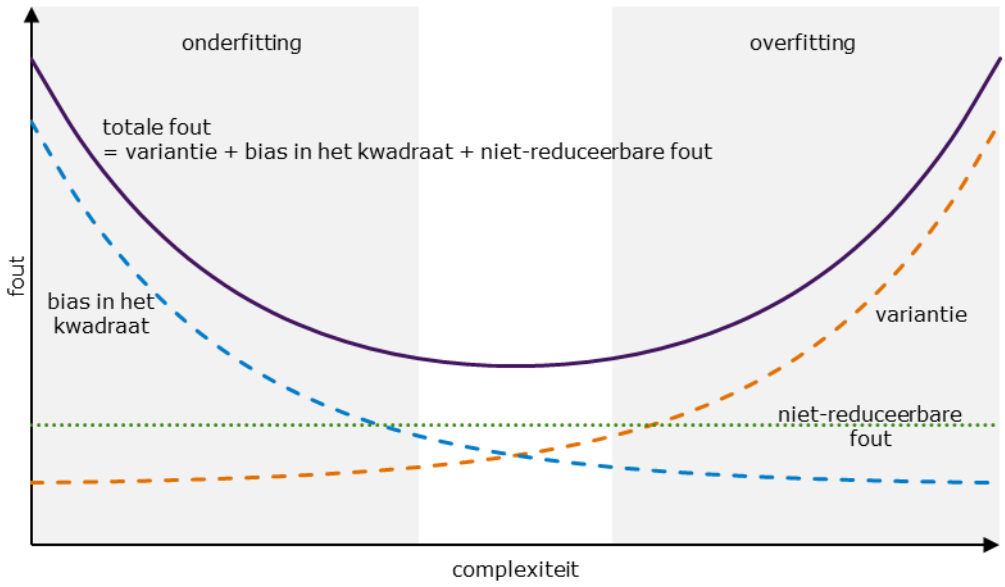
moet de afweging worden gemaakt welke voorspelfout bij voorkeur moet worden geminimaliseerd.

Bij de beoordeling van de voorspelfouten van een empirisch model zijn drie maatstaven belangrijk: 'precision', 'recall' en 'false positive rate'. Figuur 7.2 geeft aan hoe deze maatstaven worden berekend. Precision is het aantal true positives gedeeld door het aan voorspelde positives, recall is het aantal true positives gedeeld door het aantal geobserveerde positives en de false positive rate (FPR) is het aantal false positives gedeeld door het aantal geobserveerde negatives. Op het gebruik van deze maatstaven wordt nader ingegaan in de volgende paragraaf.

7.2.2 Machine learning benadering

In een machine learning model willen we zowel de bias als de variantie zo laag mogelijk houden. Maar deze grootheden bijten elkaar (zie figuur 7.2). In de praktijk blijken eenvoudige empirische modellen vaak een hoge bias en lage variantie te hebben, terwijl complexe empirische modellen vaak een hoge variantie en een lage bias hebben. Het doel van een machine learning algoritme is om de optimale combinatie van bias en variantie te vinden zodat de voorspelfout wordt geminimaliseerd. Dit wordt de 'bias-variance trade-off' genoemd. Samen met de niet-reduceerbare fout ('irreducible error'), die het gevolg is van ruis in de waarnemingen, vormen de bias en de variantie de generalisatiefout van een lerend algoritme.

Figuur 7.2 Bias-variance trade-off



Om de kwaliteit van een empirisch model te beoordelen wordt in machine learning bij het gebruik van een regressie-algoritme vrijwel altijd de gemiddelde kwadratische voorspelfout ('Mean Squared Error', MSE) berekend omdat de MSE theoretisch uitgesplitst kan worden in bias, variantie en niet-reduceerbare fout en daarmee dus de 'bias-variance trade-off' het beste meet. Maar in de praktijk kunnen de drie elementen niet apart worden berekend omdat daarvoor kennis van het werkelijke datagenererende proces (DGP) is vereist, wat meestal niet het geval is, tenzij men met kunstmatige data werkt. Een andere reden om naar de MSE te kijken is dat de

waarde van MSE alleen wordt bepaald door de data zelf en niet door de structuur van het model of de keuze van het algoritme. Daardoor kan de MSE ook gebruikt worden om verschillende type modellen en/of algoritmes met elkaar te vergelijken. In machine learning wordt doorgaans niet getoetst op problemen in de data zoals bijvoorbeeld autocorrelatie en heteroskedasticiteit (zie paragraaf 3.1).

Om een empirisch model gemaakt met een classificatie-algoritme te beoordelen, wordt gekeken naar de zogenoemde 'Receiving Operating Characteristic' (ROC)-curve. De ROC-curve zet de FPR af (op de x-as) tegen de recall (op de y-as) voor verschillende drempelwaardes. De drempelwaarde bepaalt of een voorspelde waarde als positive of negative wordt geclassificeerd (zie ook paragraaf 4.5.4). Bijvoorbeeld als de drempelwaarde 0,5 is, dan wordt een voorspelde kans tussen 0 en 0,49999 geclassificeerd als negative en een voorspelde kans tussen 0,5 en 1 als positive. Hoe groter het gebied onder de ROC-curve ('area under the curve', AUC-ROC), des te beter de prestaties van het model, onder voorwaarde dat false positives net zo erg zijn als false negatives. Als false negatives erger zijn dan false positives dan kan beter gekeken worden naar de precision-recall (PR)-curve. De PR-curve zet de recall af (op de x-as) tegen de precision (op de y-as) voor verschillende drempelwaardes. Hoe groter het gebied onder de PR-curve (AUC-PR), des te beter de prestaties van het model.

7.2.3 Econometrische benadering

In de econometrie speelt de bias-variance trade-off nauwelijks een rol, omdat de theorie zegt dat je altijd een unbiased schattingsmethode met de laagste variantie moet kiezen. Als deze schattingsmethode tevens een lineaire functie van de (getransformeerde) data is, dan spreken we van een 'best linear unbiased estimator' (BLUE).

Om de kwaliteit van een empirisch model te beoordelen wordt in de econometrische benadering zowel naar de fit en statistische toetsen als naar de voorspelfout (zoals de MSE) gekeken. Voor de beste fit wordt gekeken naar informatiecriteria gebaseerd op de waarde van de aannemelijkheidsfunctie en het aantal parameters in het model. Er zijn meerdere varianten maar alle criteria zoeken een balans tussen hoe goed het empirische model op de data past en het aantal geschatte parameters dat daarvoor nodig is, d.w.z de complexiteit van het empirische model. Daarnaast wordt met behulp van statistische toetsen getoetst op de afwezigheid van diverse problemen in de data zoals bijvoorbeeld autocorrelatie en heteroskedasticiteit (zie ook paragraaf 3.1).

Drie noodzakelijke vereisten voor vergelijking van empirische modellen op basis van informatiecriteria zijn dat de endogene variabele en de steekproef in alle empirische modellen hetzelfde zijn en dat de empirische modellen tot dezelfde klasse behoren. Deze informatiecriteria kunnen dus niet worden gebruikt om een empirisch model voor een specifieke endogene variabele te vergelijken met een empirisch model voor een getransformeerde versie van dezelfde endogene variabele (bijv. natuurlijke logaritme). Evenmin kunnen ze worden gebruikt om een empirisch model te vergelijken dat is geschat met gegevens van 2000 tot en met 2020 met een empirisch model dat is geschat met gegevens van 1990 tot en met 2020. Ook kunnen ze niet worden gebruikt om te vergelijken tussen de verschillende type empirische modellen of empirische modellen die het resultaat zijn van verschillende algoritmes. De enige manier om de verschillende type empirische modellen en/of algoritmes met elkaar te vergelijken, is om gebruik te maken van een voorspelmaatstaf, zoals de MSE. Maar omdat de 'bias-

variance trade off' een minder grote rol speelt in de econometrie zal vaker gekozen worden voor een andere voorspelmaatstaf, zoals bijvoorbeeld de gemiddelde absolute procentuele voorspelfout of de gemiddelde absolute geschaalde fout.

7.3 Implicaties voor PMJ

Met name op het gebied van machine learning hebben de afgelopen jaren veel ontwikkelingen plaatsgevonden (zie o.a. Bishop, 2006). Het PMJ is opgesteld vanuit de econometrische benadering. De vraag is of de machine learning benadering nuttig is voor het PMJ. Een andere vraag is of, ongeacht de benadering, andere algoritmes meer te bieden hebben.

7.3.1 Welke benadering?

Beide benaderingen hebben voor- en nadelen. Het doel van een machine learning algoritme is de voorspelfout van niet eerder geobserveerde gegevens te minimaliseren. De voorspelfout is het verschil tussen de prognoses gemaakt met het empirisch model en de data in de testset. De fit van het empirische model, dat wil zeggen hoe goed het empirisch model aansluit op de geobserveerde data in de trainingset, is minder relevant.[7] Ook richt machine learning zich voornamelijk op het voorspellen van reeds geobserveerde waarden/categorieën en niet op alle potentiële waarden/categorieën. Dit impliceert echter wel dat voor prognoses buiten de steekproef de niet eerder geobserveerde gegevens moeten lijken op de data waarmee het empirische model getraind is (zie box 7.1 voor een voorbeeld). Dit kan een probleem zijn bij prognoses voor de (middel)lange termijn. Naarmate de periode die voorspeld moet worden verder weg ligt van de periode waarop het empirische model is getraind, zullen de toekomstige data minder lijken op de data uit het verleden. Daardoor zijn machine learning modellen doorgaans minder geschikt voor (middel)lange termijn prognoses en meer voor classificaties (nowcasting) en zeer korte termijn prognoses (een paar maanden vooruit). In de econometrische benadering wordt zowel naar de fit en statistische toetsen als naar de voorspelfout gekeken. De modelbouwer bepaalt uiteindelijk wat de doorslag geeft, afhankelijk van het doel van de analyse. Maar omdat een econometrisch model vanuit de theorie wordt opgebouwd, is het minder gevoelig voor fluctuaties in de nog niet geobserveerde data en meer geschikt voor (middel)lange termijn prognoses. Maar als de theorie (achteraf) niet correct blijkt te zijn, kan dit leiden tot misspecificatie van het model en daarmee gepaard gaande voorspelfouten.

[7] Zie paragraaf 4.1.2 voor de definities van trainingset, validatieset en testset.

Box 7.1 Voorspellen van niet eerder geobserveerde waarden/categorieën in econometrie versus machine learning

Bij boetes voor te hard rijden loopt het boetebedrag op met de mate waarin de maximumsnelheid is overschreden. Stel we willen bepalen hoeveel snelheidsovertredingen worden begaan als het boetebedrag erg hoog wordt en we hebben tot nu toe de volgende aantallen waargenomen:

Snelheids-overtreding	≤5 km	>5, ≤ 10 km	>10, ≤15 km	>15, ≤20 km	>20, ≤25 km	>25, ≤30 km	>30 km
Boete	€ 35	€ 74	€ 139	€ 197	€ 267	€ 346	>€ 346
Aantal	5.000.000	2.500.000	1.040.000	440.000	210.000	105.000	onbekend

In de machine learning benadering wordt alleen naar de data gekeken. Omdat de data geen bestuurders bevat die meer dan 30 km/u te hard hebben gereden, zal deze categorie niet worden meegenomen in het empirische model en kunnen de aantallen hiervoor niet worden voorspeld. Maar de econometrische benadering gaat uit van de theorie. Stel dat op basis van de wetenschappelijke literatuur bekend is dat er een exponentieel verband bestaat tussen de hoogte van het boetebedrag en de snelheidsovertreding. Dan kan dit worden meegenomen in het theoretische model waardoor toch een voorspelling kan worden gemaakt voor het aantal bestuurders dat meer dan 30 km/u te hard rijdt.

Een ander verschil tussen econometrie en machine learning zit in de selectie van de exogene variabelen.[8] In de econometrie worden de exogene variabelen doorgaans geselecteerd op basis van theorie, correctheid van het teken van het geschatte effect en statistische significantie van het effect. Het toetsen van hypotheses is zeer belangrijk in een econometrische benadering, maar bij machine learning speelt dit nauwelijks een rol. In een machine learning algoritme worden de exogene variabelen geselecteerd op basis van de voorspellende waarde: als een exogene variabele de voorspellende prestaties van het empirische model verbetert, dan wordt deze meegenomen en anders weggelaten. Of de tekens van de geschatte wegingsfactoren correct zijn en of deze uiteindelijk statistisch significant zijn, is minder belangrijk. Merk op dat voor elke exogene variabele die wordt meegenomen in het model, de complexiteit van het model zal toenemen (zie ook het voorbeeld in box 7.2).

Box 7.2 Variabelen selectie in econometrie versus machine learning

Stel we willen een prognose maken van het aantal zaken dat instroomt bij het OM. De exogene variabelen die we beschikbaar hebben zijn het aantal verdachten, het aantal politie-agenten en het aantal werklozen. Als na schatting blijkt dat het aantal werklozen geen statistisch significant effect heeft op de instroom van het OM, zal deze variabele in een econometrisch model worden weglaten. Volgens het machine learning principe nemen we alle beschikbare informatie mee ook al heeft het geen statistisch significant effect op de instroom OM, zolang het empirische model maar goed voorspelt. Als na schatting van de parameters van het theoretische model blijkt dat het teken van verdachten negatief is, zullen we het aantal verdachten in het econometrisch model toch weglaten, want het is niet logisch dat de instroom OM toeneemt als het aantal verdachten afneemt. In de machine learning benadering is het teken van het aantal verdachten niet relevant, zolang het empirische model maar goed voorspelt.

[8] Zie de inleiding van hoofdstuk 3 voor de definitie van exogene variabelen.

Op het eerste gezicht lijkt de machine learning benadering aantrekkelijk voor het PMJ. Het doel van het PMJ is de best mogelijke prognose van de behoefte in justitiële ketens te maken en de machine learning benadering is volledig daarop gericht. Maar als aan de bovengenoemde randvoorwaarden moet worden voldaan, lopen we tegen de grenzen van machine learning aan. Ten eerste speelt de tijdscomponent een belangrijke rol in het PMJ maar de machine learning benadering heeft hiermee moeite. Zo vormt bootstrapping op tijdreeksen een uitdaging en is kruisvalidatie op tijdreeksen zeer lastig, omdat er dan op willekeurige plekken gaten in de tijdreeks ontstaan en de testset mogelijkerwijs niet geschikt is om het model gebouwd op de trainingsset te evalueren. Maar bootstrapping en kruisvalidatie zijn wel hele belangrijke componenten van de machine learning benadering. Ten tweede is in de machine learning benadering de gemiddelde kwadratische fout berekend op de testset bepalend voor de keuze van het model. Impliciet wordt daarbij aangenomen dat de gegevens in de testset onder dezelfde externe omstandigheden tot stand zijn gekomen als de gegevens in de trainingset. Als dat namelijk niet het geval zou zijn, zouden de externe omstandigheden de uitkomsten van het model kunnen beïnvloeden. Bij tijdreeksen is dit zeker niet het geval. Bijvoorbeeld, misdrijven geregistreerd in 2020 zijn onder andere economische, politieke en maatschappelijke omstandigheden tot stand gekomen dan misdrijven geregistreerd in 2015. Deze omstandigheden kunnen invloed hebben op de gegevens zelf. Denk hierbij bijvoorbeeld aan COVID-19. Als de trainingset gegevens tot en met 2019 bevat en de testset gegevens vanaf 2020, dan is deze testset niet geschikt om het model gebouwd op deze trainingset te evalueren.

Omdat kruisvalidatie en de gemiddelde kwadratische fout in de econometrische benadering een minder belangrijke rol spelen, zijn de randvoorwaarden van ketenconsistentie en tijdscomponent daarin makkelijker te realiseren. Maar ook de econometrische benadering heeft beperkingen. De econometrische benadering streeft naar een unbiased schatter.[9] Dit sluit algoritmes die zich vooral richten op vermindering van de variantie dus min of meer uit. Daarnaast spelen in de econometrische benadering theorie en statistische fit van het model een belangrijke rol. Dit maakt de prognoses inhoudelijk beter uitlegbaar, maar kan ook leiden tot suboptimale prognoses. Want een model dat goed fit, voorspelt niet noodzakelijkerwijs goed en vice versa.

Voor beide benaderingen is de conclusie dat de randvoorwaarde van inhoudelijk uitlegbare prognoses in veel gevallen zal botsen met het doel van de best mogelijke prognoses. Tot op zekere hoogte is er een middenweg mogelijk. Tot voor kort waren machine learning modellen vooral gebaseerd op correlaties en niet zozeer op causale verbanden. Een recente ontwikkeling is de groeiende aandacht voor causale en uitlegbare machine learning technieken (Bareinboim & Pearl, 2016; Beckers, 2022), hetgeen ook weer implicaties heeft voor de econometrische benadering (Hünermund & Bareinboim, 2023). Daarmee groeien de machine learning benadering en de econometrische benadering naar elkaar toe. Dit lijkt ook de meest belovende ontwikkelingsrichting voor het PMJ te zijn, maar de zogenoemde 'explainable artificial intelligence' (XAI) staat nog in de kinderschoenen. Uiteindelijk zal de eindgebruiker moeten aangeven welke randvoorwaarden de doorslag moeten geven en de nadelen die daarmee gepaard gaan moeten accepteren.

[9] Zie paragraaf 4.1.3 voor de definitie van bias.

7.3.2 Welk algoritme?

Kijkend naar de randvoorwaarden vallen een aantal algoritmes bij voorbaat af. Sommige algoritmes, zoals neurale netwerken, SVM/SVR, robuuste en bayesiaanse lineaire regressie, zijn te rekenintensief en kunnen daardoor niet binnen de doorlooptijd worden geactualiseerd. Algoritmes zoals neurale netwerken, discriminantenanalyse, naïeve Bayes classificatie, ETS, bayesiaanse regressie vallen ook af omdat ze niet voldoen aan de voorwaarden van ketenconsistentie en/of inhoudelijke uitlegbaarheid. Ook IV, robuuste, afgeknotte of gecensureerde regressie vallen af omdat deze algoritmen vooral bedoeld zijn voor specifiek type data of dataproblemen, die binnen het PMJ niet of nauwelijks aan de orde zijn. Gegeven de randvoorwaarden bieden met name elastic net regularisatie, survivalanalyse, ARIMA(X), ECM, logistische regressie, beslisbomen, KNN, ensemble averaging en bagging nieuwe mogelijkheden.

Het PMJ valt grofweg in vier delen uiteen: het begin van de keten, dat wil zeggen instroom in één van de justitiële ketens vanuit de maatschappij, de uitstroom bij een ketenpartner, dat wil zeggen de wijze van afdoening van ingestroomde zaken, de doorstroom door de keten, dat wil zeggen de overdracht van dossiers van de ene naar de andere justitiële ketenpartner, en de duur van bepaalde type sancties. Momenteel wordt in alle delen min of meer dezelfde technieken toegepast. Dit is echter niet noodzakelijk. Daarom moet voor alle delen apart worden gekeken wat mogelijke alternatieven zijn.

Voor het begin van de keten is momenteel alleen geaggregeerde data beschikbaar. Hierdoor staan voor het begin van de keten slechts een beperkt aantal technieken ter beschikking: gewone lineaire regressie, tijdreeksanalyse (beide reeds in het huidige PMJ), ridge regressie (evt. i.c.m. lasso), KNN, bagging en ensemble averaging. Het is onwaarschijnlijk dat in de nabije toekomst micro-data beschikbaar komt. Voor het doorstroomgedeelte van het PMJ zijn de mogelijkheden ook beperkt, omdat de micro-data van de ene ketenpartner vaak niet goed aansluiten op de micro-data van de andere ketenpartner. Hierdoor moet vaak worden teruggevallen op geaggregeerde data. Hiervoor geldt hetzelfde als voor het begin van de keten.

Maar voor het uitstroomgedeelte van het PMJ zijn er, afhankelijk van de ketenpartner, soms wel meer alternatieven beschikbaar. Voor vervolging en berechting in eerste aanleg zijn micro-data beschikbaar. Afgezien van de eerdergenoemde algoritmes, behoren (een random forest van) beslisbomen en (multinomiale) logit modellen hier ook tot de mogelijkheden. Ook voor de tenuitvoerlegging van intramurale sancties zijn microdata beschikbaar. Daarop zou een survivalanalyse kunnen worden toegepast. Het voordeel van technieken op microdata is dat ze goed zijn in het vinden van causale relaties. De gevonden causale verbanden kunnen worden geïmporteerd in een model op een hoger aggregatieniveau. Het voordeel hiervan is dat de analyse op microdata niet per se elk jaar binnen het tijdsbestek van zes weken hoeft te worden uitgevoerd. Een nadeel is dat deze technieken vaak niet goed kunnen omgaan met de tijdscomponent. De gevonden resultaten moeten dan naar de toekomst toe als constant worden beschouwd.

Ook gegevens over de duur van een sanctie zijn grotendeels alleen op een hoog aggregatieniveau beschikbaar. Maar in enkele gevallen zijn wel microdata beschikbaar, bijvoorbeeld de duur van de uit te zitten vrijheidsstraf bij DJI. Voor een groot deel van de gedetineerden is de duur vooraf bekend omdat dit bepaald is door de rechter. In

dat geval hoeft er geen schatting te worden gemaakt maar kan met een vervalkalender worden berekend wanneer mensen uitstromen. Maar in een aantal gevallen is de einddatum vooraf niet bekend, bv. bij een tbs of pij-maatregel of voorlopige hechtenis. Hiervoor zou een schatting van een survivalfunctie op microdata uitkomst kunnen bieden.

Bij het gebruik van microdata moet wel een kanttekening worden geplaatst. Wettelijk mogen organisaties alleen die informatie verzamelen die van belang is voor hun werkzaamheden. Dat betekent dat de genoemde microbestanden relatief weinig persoonskenmerken bevatten. Het aantal persoonskenmerken zou eventueel kunnen worden vergroot door te koppelen met de Sociaal Statistisch Bestanden (SSB) van het Centraal Bureau voor de Statistiek. Maar daaraan kleven wel een aantal bezwaren. Ten eerste is het de vraag of de koppeling jaarlijks binnen de gestelde termijnen kan worden gerealiseerd, aangezien het inhoudelijk en administratief gezien veel werk is en vertraging van bijvoorbeeld een week al problematisch is. Ten tweede betekent een koppeling dat gevoelige justitiële informatie bij een niet-justitiële organisatie terechtkomt en de vraag is of dit wenselijk is en of het doel deze koppeling rechtvaardigt. De privacyregels voor strafrechtelijke gegevens zijn strenger dan voor gewone persoonsgegevens. Ten derde hebben veel personen in het justitiële systeem een niet-Nederlandse nationaliteit. De vraag is of de persoonskenmerken van deze mensen in voldoende mate aanwezig zijn in het SSB. Een algoritme trainen op basis van gegevens waarin specifieke persoonskenmerken van specifieke groepen ontbreken, zou kunnen leiden tot een foutief model.

7.4 Aanbevelingen

De meest belovende ontwikkelingsrichting voor het PMJ lijkt ‘explainable artificial intelligence’ te zijn, maar dit is op dit moment nog een kennisgebied in ontwikkeling. Gegeven de aard van de data, het doel van PMJ en de randvoorwaarden leidt de voorgaande discussie tot een aantal veelbelovende algoritmes die interessant zijn om nader te onderzoeken:

- lineaire regressie met elastic net regularisatie;
- survivalanalyse op de duur van de tbs-maatregel of de voorlopige hechtenis;
- logistische regressie op beslissingen het OM of het ZM;
- random forest op beslissingen het OM of het ZM;
- k-nearest neighbours;
- ensemble averaging;
- bagging van gebootstrapte steekproeven;
- tijdreeksanalyse ten behoeve van ensemble averaging.

Grotendeels voldoen deze algoritmes aan de randvoorwaarden, maar soms zullen er concessies moeten worden gedaan. In een vervolgonderzoek zullen een aantal pilots met deze algoritmes op een beperkt aantal onderdelen van de justitiële ketens worden uitgevoerd om te kijken of deze algoritmes ook daadwerkelijk tot een hogere voorspelkwaliteit leiden.

Verder is uit een eerdere analyse van het gevangeniswezen (Moolenaar & Ter Braak, 2022) reeds naar voren gekomen dat tijdreeksanalyse op hoogfrequente bezettingsgegevens weliswaar verleidelijk is vanwege de eenvoud maar niet aan te raden vanwege seizoenseffecten, stabiliteitsissues, representativiteit van de

peilmomenten en tegen elkaar in werkende trends van instroom en gemiddelde duur. Daarom adviseerden de onderzoekers om in de toekomst aantallen en gemiddelde detentieduur indien mogelijk apart te analyseren. Uit de internationale literatuur blijkt dat dit ook gebruikelijker is.

Daarnaast is het ook belangrijk om na te denken over de wijze waarop met de prognoses en voorspelfouten wordt omgegaan. Bij elke prognose horen onzekerheidsmarges. Deze moeten worden meegewogen bij het maken van beslissingen op basis van prognoses. Ook kan worden nagedacht hoe voorspelde aantallen het best kunnen worden vertaald naar budget en/of personeel. Omdat afwijkingen tussen prognose en realiteit nooit te voorkomen zijn, is het belangrijk dat de eindgebruikers van prognoses vooraf nadenken hoe daarop te reageren. Bijvoorbeeld, hoe wil men omgaan met false positive en false negatives of te hoge en te lage voorspellingen? Wat is de economische en maatschappelijke schade? Zijn beide type voorspelfouten even erg, of is de één minder erg dan de ander? Is een organisatie flexibel genoeg ingericht om hier snel op te reageren? Dit soort vragen kunnen worden beantwoord voordat de prognoses bekend zijn, zodat protocollen kunnen worden opgesteld over hoe om te gaan met voorspelfouten.

Tot slot is het belangrijk om te realiseren dat er op het gebied van justitie en veiligheid geen absolute waarheid is. Er zijn geen wetmatigheden zoals in de exacte wetenschappen vaak wel het geval is. Dit maakt evaluatie van de prognoses achteraf lastig, want tegen welke ‘waarheid’ moeten de prognoses worden afgezet? Er kan zelfs sprake zijn van een ‘self-denying prophecy’: als beleidsmakers niet willen dat prognoses bewaarheid worden, dan kunnen ze nieuw beleid maken om ervoor te zorgen dat het niet gebeurt. Het gevolg daarvan is dat achteraf gezien de prognoses niet zijn uitgekomen, wat dan vaak ten onrechte als een fout van het model wordt aangemerkt.

Summary

Forecasting for the justice system

An exploration of different algorithms

Policymakers would like more insight into the demand for and (social) costs of crime, law enforcement and conflict resolution. It is therefore important to understand future trends in this area so that the best possible policy and financial decisions can be made. Forecasting models can be used for this purpose. Many years ago, the Forecasting Model for the Justice System (PMJ) was developed for the ministry of Justice and Security. This report examines the feasibility and usefulness of modernizing the PMJ using new developments in data and algorithms.

Forecasting model for the judicial system

Currently, forecasts about recorded crime, suspects and everything that follows, and conflict resolution are made with the PMJ. This model includes virtually the entire judicial system, including (police) investigation, prosecution and sentencing, sanctions, prisons, probation, subsidized legal aid in criminal cases and victim care. In addition, the model also includes (legal aid in) civil justice, administrative justice, legal aid in civil and administrative cases, and the detention of illegal immigrants. The judicial system can be viewed as a network. The PMJ uses demographic and economic forecasts to forecast the inflow at the beginning of the network, such as recorded crime. This forecast is then used to make forecasts of the inflow of cases at the Public Prosecution Service, and thus forecasts for the inflow at the courts and subsequently forecasts for the required sanction capacity. The PMJ is a combination of structural models, stock-flow models and time series models and includes approximately 6,600 equations. The parameters of the theoretical model are estimated using regression analysis on annual data. The result is an empirical model with which forecasts can be generated, which serve as a basis for a large part of the budget of the Ministry of Justice and Security.

Previous external evaluations show that the PMJ is well constructed and that users do not need a radically different model. But the current PMJ was designed in a period when the availability of microdata was limited, and a number of techniques were often theoretically known but could not be implemented due to limitations in computer technology. It is therefore useful to investigate whether there have been developments in recent years in the field of data and forecasting techniques that can provide more or different insights.

Limitations

The purpose of the PMJ is to make estimates of the capacity needs of the judicial system. The PMJ assumes that the entire capacity requirement can be financed. There are therefore no budget restrictions in the PMJ. The PMJ also takes chain effects into account. Only amounts are forecasted, such as number of cases or number of suspects

or number of sanctions. The PMJ does not include prices. The PMJ forecasts the items on which judicial organizations are financed. This varies per organization and is determined by the organizations themselves in consultation with the Ministry of Justice and Security and not by the PMJ. The PMJ follows these decisions. The PMJ does not determine what type of products or services or how organisations should be financed, but only how much should be financed, given what and how. If the method of financing is adjusted, the PMJ will be adjusted. A condition for the PMJ is that the criteria on which finances are provided, are quantifiable and measurable. Because the method of financing is a decision made outside the PMJ, this report will not discuss it further. The estimates themselves and their accuracy are also not discussed here. These can be found in other WODC-reports. The focus of this report is on models that can be used to forecast trends for (long-term) strategic purposes and not on predictive models for operational or forensic purposes, such as predictive policing or predictive sentencing.

Research question and preconditions

During the 2019 budget debate, the Minister for Legal Protection promised that he is willing to take another look at the PMJ. It was decided to take a two-track approach to revise the current PMJ. Track 1 concerns maintenance of and minor improvements and additions to the current PMJ. Track 2 concerns fundamental research into methods and techniques for better estimates. Track 2 is divided into three stages. In the first stage, an inventory was made of the needs of the end users of the PMJ. This stage was completed in 2020. The second stage examined the extent to which new developments in the field of data and techniques could be used in the PMJ. In the third stage, some promising techniques will be further developed in the form of pilots.
This report reports on the second stage. A large number of techniques were examined that could potentially be relevant for the PMJ. This means that with this selection of techniques the goal of the PMJ could in principle be achieved, namely making estimates of the capacity needs of the judicial system and forecasts for budget purposes. Techniques that are not suitable for this purpose have been ignored. The techniques have been assessed on the following aspects:

1 *Explainability of the algorithm from a non-technical point of view*. How easy is it to explain in simple terms what the algorithm does? In short, how intuitive is the algorithm?
2 *Simplicity of the algorithm*. How simple is the algorithm from a mathematical/statistical point of view?
3 *Implementability*. How much work does it take to implement the algorithm?
4 *Domain knowledge*. Is it possible to introduce domain knowledge into the algorithm?
5 *Network consistency*. Is it possible to use an algorithm to create a network-consistent model, that is, a model in which the outflow of one partner constitutes inflow for the next partner?
6 *Time component*. Is it possible to include a time component in the algorithm? That is, can the algorithm dynamically make a forecast for the (medium) long term or should some parts be assumed constant?
7 *Dealing with noise in the data*. Can the algorithm deal with noise in the data or does the quality of the data have to be very high?
8 *Privacy*. To what extent are micro-data necessary or can the algorithm also be applied to aggregated data? And if micro data is chosen, can the results be

aggregated in such a way that they are suitable for further processing in aggregated model?
9 *Computing time*. How much calculation time does it take to make forecasts?
10 *Explainability of the forecasts from a non-technical point of view*. Are the forecasts logical and can be explained in simple terms? Can the forecasts be traced back to specific input variables or is everything connected?
11 *Fairness*. To what extent can unfortunate choices or decision rules lead to the algorithm unintentionally becoming discriminatory?

This report will look at alternative forecasting methods, what kind of data is required for this and to what extent they could be applied without negating the advantages of the current PMJ, in particular network consistency. Therefore, any alternative algorithms must meet a number of prerequisites:

- The empirical model must be network consistent. The forecast of the output of one network partner must affect the forecast of the inflow of the subsequent network partner.
- It must be possible to predict seven years ahead, i.e. the budget horizon plus the years between the last known data point and the first budget year.
- The forecasts must be explainable from a non-technical point of view. Policymakers want to be able to understand why the forecasts are the way they are. In practice, this means that they must be traceable to specific input variables and that the estimated relationship must contain a certain degree of logic.
- Due to the planning of the budget process, the parameters of the model must be updated annually by mid-November. Because some data is only available at the end of September, this means in practice that the update must take place within a period of approximately six weeks.
- The chosen algorithm must be fair. Choices or decision rules may not unintentionally lead to the algorithm becoming discriminatory in nature.

Alternative methods

The examined techniques originate from the fields of machine learning and econometrics. Although there is a large overlap between the techniques used in econometrics and machine learning, econometrics works more from theory and machine learning is more data-driven. Roughly speaking, the alternative methods fall into four categories: alternative ways of estimating the parameters of the current PMJ, a different specification of (parts of) the model, methods that relate to greater use of the datasets that are used for estimating and testing the parameters of the model and combining methods or samples.

Alternative ways of estimating the parameters of the current PMJ

Linear regression is a simple and therefore frequently used class of algorithms for linear or linearised models. So far, the equations in the PMJ are mostly estimated with ordinary least squares. However, an algorithm that may be interesting for the PMJ is linear regression with elastic net regularization and in particular ridge regression. This method imposes a penalty on excessive complexity (i.e. the number of exogenous variables) of the model. If the data contains any significant problems, such as outliers, poorly measured or correlated exogenous variables, truncated or censored data, and

small or skewed samples a modified linear algorithm can be applied. However, often these problems can also be solved in a simpler way.

Other specification of (parts of) the PMJ

If the variable to be forecasted is an amount of some sort, both linear and non-linear regression can be used, such as linear or non-linear time series analysis or survival analysis (non-linear). Time series analysis is widely used for forecasting because it is relatively easy to implement and is integrated in many software packages. Linear time series analysis is already used to a limited extent in the current PMJ. However, the data must meet several conditions. A disadvantage is that time series analyses are mainly suitable for short-term forecasts because time series models tend to revert to the mean of the process in the long term. Other forms of non-linear regression that may be of interest to PMJ are duration data algorithms. In particular, a (semi-) parametric survival analysis of duration data seems promising for those parts of the justice field for which the duration is not known in advance, such as pre-trial detention or indefinite detention in a psychiatric prison hospital.

For the forecasts of certain choices there are algorithms such as discriminant analysis, naive Bayes algorithm or logistic regression. For the PMJ, logistic regression seems to be a logical option: there are many choices within the judicial system. Should a suspect be held in remand custody? Should a suspect be prosecuted and/or tried? What type of sanction should be imposed? These are typical choices that can be predicted with a logistic regression. Nevertheless, there are also limitations. The purpose of the PMJ is to predict seven years ahead. It is difficult to include the time aspect in a logistic regression. Furthermore, logistic regression is applied to microdata. The question is whether the risk of privacy violations outweighs increased insight into the forecasts. Discriminant analysis is a simple, computationally efficient algorithm, especially suitable for small samples with a limited number of exogenous variables. The naive Bayes algorithm is also technically easy, but the results are less intuitive if one has no knowledge of or affinity with probability distributions. This makes both algorithms less useful for the PMJ.

There are also algorithms that can predict both amounts and choices, such as k-nearest neighbours, decision trees, support vector machines and neural networks. These non-parametric algorithms have the advantage that very little assumptions about the data are made in advance. The advantages and disadvantages differ per algorithm. All things considered, k-nearest neighbours and decision trees could be a good addition to the PMJ, where the main disadvantages of not being able to trace back to specific background factors or the absence of the time component, respectively, will have to be weighed against the advantages. Support vector machines and neural networks are less likely candidates mainly because the calculation time of these algorithms is large, and the updating of the models cannot be achieved in a satisfactory manner within the available time (approximately six weeks). Another disadvantage, particularly with neural networks, is that everything is interconnected, and forecasts cannot be explained from a non-technical point of view.

Greater utilization of the dataset

There are several methods which make broader use of the current sample. One way is to improve the validation errors or the standard errors and the confidence intervals of the estimated parameters. By means of bootstrapping or cross-validation, multiple

subsets can be drawn from the same sample, with which the same model is estimated each time. These methods are used to compare different theoretical models and/or algorithms with each other in terms of predictive power. Bootstrapping in particular, can be interesting, especially in combination with alternative algorithms. But with a large number of models, algorithms and/or observations, it can be very computationally intensive. Moreover, both methods are somewhat difficult to apply to time series. Also, because the PMJ is usually not assessed solely on predictive power, but on several criteria, including theoretical correctness and explainability from a non-technical point of view, especially with cross-validation the marginal advantage of better predictive criteria and knowledge about the distribution of the parameters does not seem to outweigh the disadvantage of greater computational intensity.

Another way of making broader use of the current sample is to include combinations of exogenous variables in the model instead of individual exogenous variables, for example by means of principal component analysis or factor analysis. The major disadvantage is that the resulting forecasts cannot be traced back to specific background factors and can therefore not be explained from a non-technical point of view. Therefore, the application of principal components analysis or factor analysis is less desirable for PMJ.

Combining methods or sampling

There are several ways in which different forecasts of the same variable can be combined. Both bagging and boosting combine forecasts from the same algorithm but different samples. Bagging, or bootstrap aggregating, could be interesting for the PMJ, because it is relatively easy to implement and the model itself does not become too complex, so that the results are still explainable. Bagging is also an essential part of the random forest algorithm. Random forest is a forest of decision trees, where each decision tree is built on a different bootstrapped sample of observations and/or a different random selection of the exogenous variables. Boosting is less interesting because it requires very high data quality without outliers and noise and it is computationally intensive, meaning that the forecasts cannot be updated within the available time.

Ensemble averaging and stacking allow forecasts from different algorithms to be combined on the same sample. With ensemble averaging, an average is calculated, while with stacking a metamodel is formulated. Ensemble averaging is a promising technique for the PMJ, because it is relatively easy to implement, and different algorithms are often tried out in the test phase anyway. Until now, one model was ultimately chosen based on various criteria, although the differences in prediction quality were often minor. Alternatively, the average of the forecasts from multiple models could be calculated. For example, the inflow into the Public Prosecution Service could be predicted based on the number of suspects recorded by the police, but also by means of a time series analysis. The final forecast then becomes the (weighted) average of the outcomes of both algorithms. Stacking could be an option if the metamodel is not too complicated. But the disadvantage is that it takes a lot of calculation time and requires (relatively) a lot of data.

Conclusion and recommendations

In contrast to econometric models, machine learning models are mainly data-driven and therefore mainly based on correlations and not so much on causal relations. A recent development is the growing attention for causal and explainable machine learning techniques, the so-called 'explainable artificial intelligence' (XAI). This means that the machine learning models and the econometric models are evolving toward each other. This also seems to be the most promising development direction for the PMJ, but XAI is an area of knowledge that is still under development. Given the nature of the data, the purpose of PMJ and the prerequisites, the most promising alternative algorithms are:

- algorithm that imposes a penalty on excessive complexity of the model (linear regression with elastic net regularization);
- analysis of duration data for psychiatric prison orders or pre-trial detention (survival analysis);
- algorithm that assumes that similar characteristics of the exogenous variables lead to comparable values of the endogenous variable (k-nearest neighbours);
- algorithm for choices, such as the type of punishment to be imposed (logistic regression, decision tree);
- broader utilization of the existing data set (bagging of bootstrapped samples);
- combination of multiple decision trees through bagging (random forest);
- combination of the results of different algorithms (ensemble averaging);
- time series analysis for ensemble averaging.

These algorithms largely meet the prerequisites, but sometimes concessions will have to be made. In a follow-up study, several pilots will be carried out with these algorithms to see whether these algorithms actually lead to better forecasting performance.

Literatuur

Arrow, K. J. (1963). *Social Choice and Individual Values* (2de editie). Wiley.

Ashby, M. P. J. (2020). Initial evidence on the relationship between the coronavirus pandemic and crime in the United States. *Crime Science, 9*(1), 6. https://doi.org/10.1186/s40163-020-00117-6

Bareinboim, E., & Pearl, J. (2016). Causal inference and the data-fusion problem. *Proc Natl Acad Sci U S A, 113*(27), 7345-7352. https://doi.org/10.1073/pnas.1510507113

Barros, P. H. B. d., Baggio, H. d. S., & Baggio, I. S. (2020). The Socioeconomic Determinants of Crime in Brazil: The role of spatial spillovers and heterogeneity. *Revista brasileira de segurança pública, 14*(2), 188-209. https://doi.org/10.31060/rbsp.2020.v14.n2.1091

Becker, G. S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy, 76*(2), 169-217. https://doi.org/10.1086/259394

Beckers, S. (2022). Causal explanations and XAI. *1st Conference on Causal Learning and Reasoning. Proceedings of Machine Learning Research, 140,* 1-20.

Biermans, M., & Van Leeuwen, M. (2003). *SWOT-analyse modellen Veiligheidsketen*. SEO. Rapport nr. 699.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer.

Bomhoff, E. J., Van der Voort van der Kleij, G. T., & Sadiraj, K. (2002). *Tekort aan cellen*. NYFER.

Bont, P. F. H., Homburg, G. H. J., & Van Rij, C. (2009). *Evaluatie PMJ-systeem: van beleidsneutraal naar beleidsrijk*. Regioplan. Publicatienr. 1734.

Chua, E., & Tumibay, G. (2020). Crime Data Forecasting using Exponential Smoothing. *International Journal of Advanced Trends in Computer Science and Engineering, 9*(1.1 S I), 69-75. https://doi.org/10.30534/ijatcse/2020/1391.12020

De Poot, H., McKim, M., Dieleman, R., Bonenberg, B., & Brussee, R. (2020). *Klant en Keten. Het Prognosemodel Justitiële Ketens gezien door Gebruikers*. Nobis Policy Lab.

Deadman, D. (2003). Forecasting residential burglary. *International Journal of Forecasting, 19*(4), 567-578. https://doi.org/10.1016/s0169-2070(03)00091-8

Dhiri, S., Brand, S., Harries, R., & Price, R. (1999). *Modelling and predicting property crime trends in England and Wales*. Home Office. Research Study 198.

DSRS. (2002). *Prognoses sanctiecapaciteit. Evaluatie van het proces*. Ministerie van Justitie.

Everhardt, T., Vonk, J., & Wilms, P. (2016). *Review PMJ ramingen. Second opinion van de beleidsneutrale ramingen*. APE. Rapportnr. 1477.

Felsö, F., Scheele, D., Bremer, S., & Baarsma, B. (2006). *Evaluatie Prognosemodellen Justitiële Ketens: Civiel en Bestuur*. SEO. Rapportnr. 923.

Gardner, E. S. (2006). Exponential smoothing: The state of the art—Part II. *International Journal of Forecasting, 22*(4), 637-666. https://doi.org/10.1016/j.ijforecast.2006.03.005

Goldman, J., Hooper, R. L., & Mahaffey, J. A. (1976). Caseload Forecasting Models for Federal District Courts. *The Journal of Legal Studies, 5*(2), 201-242. https://doi.org/10.1086/467551

Gorr, W., Olligschlaeger, A., & Thompson, Y. (2003). Short-term forecasting of crime. *International Journal of Forecasting, 19*(4), 579-594. https://doi.org/10.1016/s0169-2070(03)00092-x

Goudriaan, R. (2004). *Beoordeling prognosemodel voor de veiligheidsketen*. APE.

Harries, R. (2003). Modelling and predicting recorded property crime trends in England and Wales—a retrospective. *International Journal of Forecasting, 19*(4), 557-566. https://doi.org/10.1016/s0169-2070(03)00090-6
Hendry, D. F., & Clements, M. P. (2004). Pooling of forecasts. *The Econometrics Journal, 7*(1), 1-31.
Hu, T., Zhu, X., Duan, L., & Guo, W. (2018). Urban crime prediction based on spatio-temporal Bayesian model. *PLOS ONE, 13*(10), e0206215. https://doi.org/10.1371/journal.pone.0206215
Huber, P. J. (1973). Robust Regression: Asymptotics, Conjectures and Monte Carlo. *The Annals of Statistics, 1*(5), 799–821.
Huijbregts, G. L. A. M., Van Tulder, F. P., & Moolenaar, D. E. G. (2001). *Model van justitiële jeugdvoorzieningen voor prognose van de capaciteit.* WODC. Onderzoek en beleid 192.
Hünermund, P., & Bareinboim, E. (2023). Causal Inference and Data Fusion in Econometrics. *Forthcoming in: The Econometrics Journal,* arXiv preprint arXiv:1912.09104., 1-44.
Hyndman, R. J., & Athanasopoulos, G. (2021). *Forecasting: principles and practice* (3de editie). Otexts.
KPMG/BEA. (1998). *De plausibiliteit van het prognosemodel sanctiecapaciteit*. KPMG Bureau voor Economische Argumentatie.
Kruisbergen, E.W., Haas, M., Moolenaar, D., Van Es, L., Snijders, J., Houwing, L, & Stickle, B. (2024). The pandemic as a criminological experiment: Crime in the Netherlands during 12 months of COVID-19 measures. *European Journal of Criminology.* 1-27.
Leertouwer, E. C., Van Tulder, F. P., Diephuis, B. J., Folkeringa, M., & Eshuis, R. J. J. (2005). *Prognosemodellen Justitiële Ketens: Civiel en Bestuur*. WODC. Cahier 2005-13.
Leertouwer, E. C., Van Tulder, F. P., Diephuis, B. J., Folkeringa, M., & Van Gammeren-Zoeteweij, M. (2007). *PrognoseModel Justitiële Ketens 2006: Onderdelen Civiel en Bestuur: Beschrijving van het verbetertraject 2005/2006*. WODC. Cahier 2007-11.
Lilly, R. J., Cullen, F. T., Ball, R. A., & Inciardi, J. A. e. (1995). *Criminological theory: context and consequences* (2de editie). Sage.
Lin, B.-S., MacKenzie, D. L., & Gulledge, T. R. (1986). Using ARIMA Models to Predict Prison Populations. *Journal of Quantitative Criminology, 2*(3), 251-264. https://doi.org/10.1007/BF01066529
Liu, H., & Brown, D. E. (2003). Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting, 19*(4), 603-622. https://doi.org/10.1016/s0169-2070(03)00094-3
McFadden, D. (1974). Conditional Logit Analysis of Qualitative Choice Behavior. In P. Zarembka (Ed.), *Frontiers in Econometrics* (pp. 105-142). Academic Press.
Moolenaar, D., & Choenni, S. (2021). The Impact of the COVID-19 Crisis on the Dutch Criminal Justice System. *Global Economics Science, 2*(3), 24-41. https://doi.org/10.37256/ges.232021945
Moolenaar, D. E. G., Huijbregts, G. L. A. M., & Van der Heide, W. (2004). *Prognosemodel Justitiële Ketens*. WODC. Cahier 2004-8.
Moolenaar, D. E. G., Kriege, A. G., & Diephuis, B. J. (2021). *Capaciteitsbehoefte Justitiële Ketens t/m 2026 (Demand for capacity in the Justice System until 2026)*. WODC/Rvdr. Cahier 2021-12.
Moolenaar, D. E. G., & Ter Braak, F. (2022). *Bezetting van het Gevangeniswezen*. WODC. Cahier 2022-11.

Moolenaar, D. E. G., Van Tulder, F. P., Decae, R., Smit, P. R., & Diephuis, B. (2018). *Terug naar de toekomst II. Het beroep op justitiële voorzieningen 2008-2017: raming en realisatie*. WODC/Rvdr. Cahier 2018-6.

Moolenaar, D. E. G., Van Tulder, F. P., & Van Gammeren-Zoeteweij, M. (2009). *Terug naar de toekomst. Het beroep op Justitie, 1997-2007: raming en realisatie*. WODC/Rvdr. Cahier 2009-6.

Payne, J. L., & Morgan, A. (2020). *Property crime during the COVID-19 pandemic: a comparison of recorded offence rates and dynamic forecasts (ARIMA) for March 2020 in Queensland* (SocArXiv, nr. de9nc ed.). Center for Open Science.

Payne, J. L., Morgan, A., & Piquero, A. R. (2022). COVID-19 and social distancing measures in Queensland, Australia, are associated with short-term decreases in recorded violent crime. *J Exp Criminol, 18*(1), 89-113. https://doi.org/10.1007/s11292-020-09441-y

Piquero, A. R., Riddell, J. R., Bishopp, S. A., Narvey, C., Reid, J. A., & Piquero, N. L. (2020). Staying Home, Staying Safe? A Short-Term Analysis of COVID-19 on Dallas Domestic Violence. *Am J Crim Justice, 45*(4), 601-635. https://doi.org/doi:10.1007/s12103-020-09531-7

Rousseeuw, P. J., & Yohai, V. J. (1984). Robust Regression by Means of S-Estimators. In J. Franke, W. Härdle, & D. Martin (Eds.), *Robust and Nonlinear Time Series*. Springer-Verlag.

Saridakis, G. (2003). *Violent Crime in the United States of America: A Time-Series Analysis Between 1960-2000* (Discussion Papers in Economics 03/14 ed.). University of Leicester.

Shoesmith, G. L. (2013). Space–time autoregressive models and forecasting national, regional and state crime rates. *International Journal of Forecasting, 29*(1), 191-201. https://doi.org/10.1016/j.ijforecast.2012.08.002

Smit, P., & Choenni, S. (2014). On the interaction between forecasts and policy decisions. *Proceedings of the 15th Annual International Conference on Digital Government Research* (pp. 110-117). Association for Computing Machinery.

Spapens, A. C., Hoogeveen, C. E., & Van Tits, M. (2001). *Evaluatie van het model Jukebox 2. Plausibiliteit van de variabelen en verklaringsrelaties in het model*. IVA.

Ter Braak, F., Bargh, M. S., Moolenaar, D. E. G., Choenni, S., & Tims, B. (2024). *Forecasting Algorithms: Technical annotation to Cahier 2024-4.* WODC.

Theeuwes, J. J. M., & De Winter, J. M. (1998). *Econometrische evaluatie 'Prognose sanctiecapaciteit'*. SEO. Rapport nr. 485.

Tims, B., Moolenaar, D. E. G., Kriege, A. G., & Van der Pol, B. (2023). *Capaciteitsbehoefte Justitiële Ketens t/m 2028. Beleidsneutrale ramingen*. WODC/Rvdr. Cahier 2023-5.

Van der Heide (red.), W. (2002). *Prognoses en simulaties op Justitieterrein. Seminar gehouden op 27 november 2001*. WODC.

Van der Torre, A. G. J., & Van Tulder, F. P. (2001). *Een model voor de strafrechtelijke keten (A model for the criminal justice system)*. Sociaal en Cultureel Planbureau.

Vijayarani, S., Suganya, E., & Navya, C. (2021). Crime Analysis and Prediction Using Enhanced Arima Model. *International Journal of Research Publication and Reviews, 2*(1), 257-266.

Virén, M. (2010). Modelling crime and punishment. *Applied Economics, 33*(14), 1869-1879. https://doi.org/10.1080/00036840010017677

Wang, X., Hyndman, R. J., Li, F., & Kang, Y. (2023). Forecast combinations: An over 50-year review. *International Journal of Forecasting, 39*(4), 1518-1547. https://doi.org/10.1016/j.ijforecast.2022.11.005

Witt, R., & Witte, A. (2000). Crime, Prison and Female Labor Supply. *Journal of Quantitative Criminology, 16*(1), 69-85. https://doi.org/10.1023/a:1007525527967

Yohai, V. J. (1987). High Breakdown-Point and High Efficiency Robust Estimates for Regression. *The Annals of Statistics, 15*(2), 642-656.

Bijlage 1 Programmeringsadviesgroep PMJ

De leden van de programmeringsadviesgroep PMJ zijn in alfabetische volgorde:

Prof. dr. P. Boswijk
Universiteit van Amsterdam

Ir. A. Christiaanse
Ministerie van Justitie en Veiligheid

Prof. dr. H. Elffers
Nederlands Studiecentrum Criminaliteit en Rechtshandhaving

Prof. dr. Ir. H. La Poutré
Centrum voor Wiskunde en Informatica

Drs B. Smid
Centraal Planbureau

Drs. A.G.J. Van der Torre
Sociaal Cultureel Planbureau (t/m september 2023)
Vanaf oktober 2023 op persoonlijke titel

Het Wetenschappelijk Onderzoek- en Datacentrum (WODC), Kennisinstituut voor de rechtsstaat, is een onafhankelijk kennisinstituut dat valt onder het ministerie van Justitie en Veiligheid. Het WODC draagt bij aan behoud en verbetering van de rechtsstaat via het (laten) uitvoeren van kwalitatief hoog wetenschappelijk onderzoek. En door het aanbieden van gevraagde en ongevraagde kennis, verbeterpunten en (waar mogelijk) denkrichtingen.

Meer informatie:

www.wodc.nl